Arslan Brömme, Christoph Busch, Antitza Dantcheva, Kiran Raja, Christian Rathgeb, Andreas Uhl (Eds.):
BIOSIG 2020 - 19th International Conference of the Biometrics Special Interest Group

GI-Edition

Lecture Notes in Informatics

Arslan Brömme, Christoph Busch, Antitza Dantcheva, Kiran Raja, Christian Rathgeb, Andreas Uhl (Eds.)

BIOSIG 2020

Proceedings of the 19th International Conference of the Biometrics Special Interest Group

16.–18. September 2020
International Digital Conference

306

Proceedings

GESELLSCHAFT
FÜR INFORMATIK

Arslan Brömme, Christoph Busch,
Antitza Dantcheva, Kiran Raja,
Christian Rathgeb, Andreas Uhl (Eds.)

BIOSIG 2020

Proceedings of the 19th International Conference of the Biometrics Special Interest Group

16.-18. September 2020
International Digital Conference

Gesellschaft für Informatik e.V. (GI)

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

Volume P-306
ISBN 978-3-88579-700-5
ISSN 1617-5468

Volume Editors

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

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

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

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

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

Andreas Uhl
University of Salzburg
Jakob-Haringer Str. 2, A-5020 Salzburg
Email: 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 well known forum for biometrics and security professionals from industry, science, representatives of the national governmental bodies and European institutions who are working in these areas.

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

The international program committee accepted full scientific papers strongly according to the LNI guidelines (**acceptance rate ~32%**) within a scientific double-blinded review process of at minimum five reviews per paper. All papers were formally restricted for the digital proceedings up to 12 pages for regular research contributions including an oral presentation and up to 8 pages for further conference contributions.

Furthermore, the program committee has created a program including selected contributions of strong interest (further conference contributions) for the outlined scope of this conference. All paper contributions for BIOSIG 2020 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.

16th September 2020

Arslan Brömme (*GI BIOSIG, GI e.V.)*, **Christoph Busch** *(Hochschule Darmstadt)*,
Antitza Dantcheva (*INRIA Méditerranée)*, **Kiran Raja** *(NTNU)*,
Christian Rathgeb *(Hochschule Darmstadt)*, **Andreas Uhl** *(University of Salzburg)*

Chairs

General Chair

Christoph Busch, *Hochschule Darmstadt, Germany*

Program Chairs

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

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

Christian Rathgeb, *Hochschule Darmstadt, Germany*

Andreas Uhl, *University of Salzburg, Austria*

Publication Chair

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

Publicity Chairs

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

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

Local Chairs

Alexander Nouak, *Fraunhofer IUK, Darmstadt, Germany*

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

Program Committee

Fernando Alonso-Fernandez (HaU, SE)
Harald Baier (HDA, DE)
Piotr Bilinski (Univ. Oxford, UK)
Patrick Bours (NTNU, NO)
Fadi Boutros (FHG IGD, DE)
Ralph Breithaupt (BSI, DE)
Julien Bringer (Smart Valor, FR)
Arslan Brömme (GI BIOSIG, DE)
Christoph Busch (CAST-Forum, DE)
Victor-Philipp Busch (Sybuca, DE)
Cunjian Chen (MSU, US)
Nasder Damer (FHG IGD, DE)
Antitza Dantcheva (INRIA, FR)
Maria De Marsico (SUoR, IT)
Max Fermann (BoNZ, NZ)
Olaf Henniger (FHG IGD, DE)
Heinz Hofbauer (PLUS, AT)
Norbert Jung (HBRS, DE)
Ali Khodabakhsh (NTNU, NO)
Ulrike Korte (BSI, DE)
Stan Li (CBSR, CN)
Johannes Merkle (secunet, DE)
Amir Mohammadi (Idiap, CH)
Aythami Morales (UAM, ES)
Emilio Mordini (RT, FR)
Abe Narishige (Fujitsu, JP)
Mark Nixon (UoS, UK)
Alexander Nouak (FHG IUK, DE)
Markus Nuppeney (BSI, DE)
Daile Osorio-Roig (HDA, DE)

Martin Drahansky (BUT, CZ)
Pawel Drozdowski (HDA, DE)
Matteo Ferrara (UoB, IT)
Julian Fierrez (UAM, ES)
Daniel Fischer (HDA, DE)
Lothar Fritsch (KAU, SE)
Steven Furnell (UoN, UK)
Javier Gallbaly (JRC, IT)
Sonia Garcia (TSP, FR)
Marta Gomez-Barrero (HAN, DE)
Lazaro Janier Gonzalez-Soler (HDA, DE)
Ester Gonzalez-Sosa (Nokia Bell Labs, ES)
Brian Greenough (BoNZ, NZ)
Marcel Grimmer (HDA, DE)
Patrick Grother (NIST, US)
Daniel Priesnitz (HDA, DE)
Kiran Raja (NTNU, NO)
Raghavendra Ramachandra (NTNU, NO)
Christian Rathgeb (HDA, DE)
Heiko Roßnagel (FHG IAO, DE)
Torsten Schlett (HDA, DE)
Günter Schumacher (JRC, IT)
Takashi Shinzaki (Fujitsu, JP)
Max Snijder (EAB, NL)
Luis Soares (ISCTE-IUL, PT)
Anna Stratmann (BSI, DE)
Ruben Tolosana (UAM, ES)
Andreas Uhl (PLUS, AT)
Andreas Wolf (BDR, DE)

Hosts

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

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

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

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

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

Norwegian Biometrics Laboratory (**NBL**)
http://www.nislab.no/biometrics_lab

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

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 2020 – Biometrics Special Interest Group

"**2020 International Conference of the Biometrics Special Interest Group**"
16th -18th September 2020

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

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

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

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

BIOSIG 2020 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 2020 – Regular Research Papers

BIOSIG 2020 – Further Conference Contributions

BIOSIG 2020

Regular Research Papers

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 1

The Effect of Wearing a Mask on Face Recognition Performance: an Exploratory Study

Naser Damer[12] , Jonas Henry Grebe[1], Cong Chen[1], Fadi Boutros[12],
Florian Kirchbuchner[1], Arjan Kuijper [12]

Abstract: Face recognition has become essential in our daily lives as a convenient and contactless method of accurate identity verification. Process such as identity verification at automatic border control gates or the secure login to electronic devices are increasingly dependant on such technologies. The recent COVID-19 pandemic have increased the value of hygienic and contactless identity verification. However, the pandemic led to the wide use of face masks, essential to keep the pandemic under control. The effect of wearing a mask on face recognition in a collaborative environment is currently sensitive yet understudied issue. We address that by presenting a specifically collected database containing three session, each with three different capture instructions, to simulate realistic use cases. We further study the effect of masked face probes on the behaviour of three top-performing face recognition systems, two academic solutions and one commercial off-the-shelf (COTS) system.

Keywords: Face recognition, COVID-19, masked face recognition.

1 Introduction

Given the current COVID-19 pandemic, it is essential to enable contactless and smooth running operations, especially in contact sensitive facilities like airports. Face recognition have been been praised as such an accurate and contactless mean of verifying identities. Wearing masks is essential to prevent the spread of contagious diseases and have been currently forced in public places in many countries. However, the performance, and thus the trust, of contactless identity verification through face recognition can be effected by wearing a mask.

Face occlusion have been repeatedly addressed in the scope of face detection solutions [Op16]. Moreover, developing occlusion invariant face recognition solutions has been a growing research challenge [So19]. However, most of these works address general occlusion that commonly appear in in-the-wild capture conditions, such as sunglasses and partial captures. Given the current COVID-19 pandemic, it is essential to study the specific effect of wearing face masks on the behaviour of face recognition system in a collaborative environment. Our work aims at studying this effect to enable the future development of solutions addressing accurate face recognition in such scenarios. To achieve that, we present a database that simulates a realistically variant collaborative face capture scenario. This database is a first version of an on-going data collection process that includes three session, each with three capture variations, per subject. We study the behaviour of three of

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany

the top performing face recognition solutions (one commercial and two academic) when encountering masked faces, in comparison to the typical no-mask baseline. We conclude with pointing out strong signs of negative effect on face recognition systems, showing the need to develop appropriate evaluation databases and recognition solutions.

2 Related work

Face recognition deployment faces a number of operational challenges. Many of these challenges, and thus the research efforts, are related to attacks on face recognition systems, such as presentation attacks (spoofing) [DD16], morphing attacks [Da19], or other unconventional attacks [Da18]. However, issues related to the biometric sample presentation, such as face occlusion, can also effect face recognition deployability. The detection of occluded faces is a well-studied issue in the computer vision domain. An example of that is the work of Optiz et al. [Op16] that proposed a novel grid loss targeting a more accurate detection of occluded faces. Focusing on masked faces, Ge et al. [Ge17] presented a solution to enhance the detection (not biometric recognition) of masked faces in in-the-wild scenarios. Their experiments did not focus on masks worn specifically for health protection reasons, but included other forms of face occlusions. However, their solution is relevant to face recognition as our experiments will show later that the investigated face recognition solutions fails in some cases to detect a face.

As stated, detecting occluded faces is a challenge that affect the operation of face biometric systems. However, the biometric recognition of these faces is a more dominant challenge. An example of the works addressing this challenge is that of Song et al. [So19] where they aim at enhancing face recognition for faces with general occlusions. Their approach tries to learn finding and discarding corrupted feature elements, linked to occlusions, from the recognition process. Focusing on masks, in a very recent work, Wang et al. [Wa20] presented, in a brief and undetailed work, crawled databases for face detection, recognition and simulated masked faces. The authors claim to enhance the recognition accuracy from 50% to 95% without providing information on their baseline, proposed algorithmic details, or clearly specifying the evaluation database. Given the current COVID-19 pandemic, a specifically collected database and evaluation of wearing real face mask on collaborative face recognition is necessary and is still missing.

3 The database

The goal of the collected database is to enable the study of face recognition performance on masked faces and drive future innovation in this domain. The database presented in this work is an initial version and further data collection efforts is on going. The data tries to simulate a collaborative, yet varying, scenario. Such as the situation in automatic border control gates or unlocking personal devices with face recognition, where the mask, illumination, and background can change.

Each of the participants was asked to collect the data on three different, not necessary consecutive days. We consider each of these days as one session. On each day, the participant will collect three videos, each of a minimum length of 5 seconds. All videos are collected from static (not hand held) webcams and the users were asked to simulate a login scenario by looking at the capture device. The images were all captured indoors, each at their residence during home-office. The capture was performed during the day (day-light) and the participants were asked to remove eyeglasses only when the frame is considered very thick. No other restrictions were imposed, such as background or mask type and its consistency over days, to simulate realistic scenarios. The three videos captured each day were as follows: 1) Face with no mask and no additional electric illumination, this will be noted as baseline (BL). 2) Face with mask on and no additional electric illumination, this will be noted as mask one (M1). 3) Face with mask on and the existing electric light in the room is turned on, this will be noted as mask two (M2). The M2 is considered to study the unknown effect of illumination variation in the case of masked face recognition, given that the mask might result in different reflection and shadow patterns.

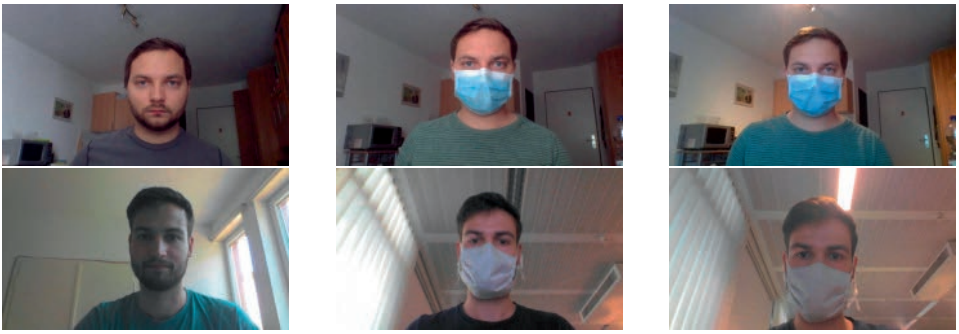
The first session (day) is considered as the reference data (R), resulting in the baseline reference (BLR), the mask one reference (M1R), and mask two reference (M2R). The second and third sessions (days) were considered as probe data (P) and they result in the baseline probe (BLP), the mask one probe (M1P), and mask two probe (M2P), and the joint probe data from M1P and M2P referred to as M12P. From each captured video, the first second was neglected to avoid any biases related to the user interaction with the capture device. From the following three seconds, 10 frames were selected with 9 frames gap between them, as all videos are recorded at 30 frames per second. The total number of participant at this first version of the database is 24, and they all participated in all sessions. Given the number of sessions, participants and the considered frames from each video, Table 1 provide an overview on the database structure. Samples of the database are shown in Figure 1.

Session	Session 1: References			Session 2 and 3: Probes			
Data split	BLR	M1R	M2R	BLP	M1P	M2P	M12P
Illumination	No	No	Yes	No	No	Yes	Both
Number of Captures	240	240	240	480	480	480	960

Tab. 1: An overview of the database structure.

4 Face recognition

To provide a wide view on the effect of wearing a mask on face recognition performance, we analyse the performance of three face recognition algorithms. Two of these algorithms are of the top performing academic approaches, namely the ArcFace [De19] and SphereFace [Li17]. The third algorithm is a COTS algorithm from the vendor Neurotechnology [Ne]. In the following, this section provides more details on these algorithms.



(a) BL (b) M1 (c) M2

Fig. 1: Samples of the collected database from the three capture types (BL, M1, and M2)

SphereFace: We chose SphereFace as it achieved competitive verification accuracy on Labeled Face in the Wild (LFW) [Hu07] 99.42% and Youtube Faces (YTF) [WHM11] 95.0% using 64-CNN layers trained on CASIA-WebFace dataset [Yi14]. SphereFace is trained using angular Softmax loss function (A-Softmax). The key idea behind A-Softmax loss is to learn discriminative features from the face image by formulating the Softmax as angular computation between the embedded features vector X and their weights W.

ArcFace: ArcFace achieved state-of-the-art performance of several face recognition benchmarks such as LFW 99.83% and YTF 98.02%. ArcFace introduced Additive Angular Margin loss (ArcFace) to enhance the discriminative power of the face recognition model. We employed ArcFace based on ReseNet-100 [He16] architecture pretrained on refined version of MS-Celeb-1M dataset [Gu16] (MS1MV2).

COTS: We used the MegaMatcher 11.2 SDK [Ne] from the vendor Neurotechnology. We chose this COTS product as Neurotechnology achieved one of the best performances in the recent NIST report addressing the performance of vendor face verification products [GP20]. The face quality threshold was set to zero to minimize neglecting masked faces. The full processes of detecting, aligning, feature extraction, and matching are part of the COTS and thus we are not able to provide their algorithmic details. Matching two faces by the COTS produces a similarity score.

For the ArcFace [De19] and SphereFace [Li17], the Multi-task Cascaded Convolutional Networks (MTCNN) [Zh16] solution is used, as recommended in [Li17], to detect (crop) and align (affine transformation) the face. Both network process the input aligned and cropped image and produce a feature vector of the size 512. To compare two faces, a distance is calculated between their respective feature vectors. This is calculated as Euclidean distance for ArcFace features, as recommended in [De19], and as Cosine distance for SphereFace features, as recommended in [Li17]. The Euclidean distance (dissimilarity) is complemented to show a similarity score and the Cosine distance shows similarity score by default.

5 Experimental setup

To baseline the performance, we evaluate the face verification performance without masks. This is done by N:N comparison of the data splits BLR and BLP (BLR-BLP). To measure the performance when wearing a mask, we perform an N:N comparison between the data splits BLR and M1P (BLR-M1P). To evaluate any induced performance change by having an additional illumination (room light) when wearing a mask, we perform an N:N comparison between the data splits BLR and M2P (BLR-M2P). To measure the overall performance including both considered illumination, we perform an N:N comparison between the data splits BLR and M12P (BLR-M12P). These four experiments are used to evaluate each of the three considered face recognition solutions.

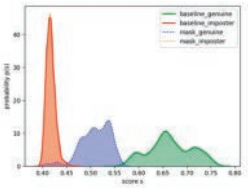
To study the effect of wearing a mask on the recognition performance, we plot the genuine and imposter distributions of the BLR-BLP (baseline) comparisons along with the genuine and imposter score distributions of the BLR-(M1P or M2P or M12P) (mask). This allows analysing the shifts in the distributions induced by wearing a mask. We also report the mean of the genuine scores (G-mean) and mean of imposter scores (I-mean) for each experiment, to get a quantitative measure of the comparison scores shifts.

Based on the standard ISO/IEC 19795-1 [Ma06], we also enrich our performance study by a number of verification performance metrics. As the face mask induces a strong appearance change on the face, face detection might be challenging. Therefore, we report the failure to extract rate (FTX) for each experiment. FTX is proportion of failures of the feature extraction process to generate a template from the captures sample. Besides reporting the FTX, and only for the samples where a template can be created, we report algorithmic verification performance metrics. These metrics include the general Equal Error Rate (EER), which is defined as the common value of false mathc rate (FMR) and false non match rate (FNMR) at the decision threshold where they are identical. We also show the algorithmic verification performance by listing the FNMR at different operation points by presenting the achieved FMR100, FMR1000, and ZeroFMR, which are the lowest FNMR for an FMR $\leq$1.0%, $\leq$0.1%, and $\leq$0%, respectively. To provide an algorithmic verification performance illustration on the complete range of operation points, we plots the receiver operating characteristic (ROC) curves for all the experimental setups, for each of the investigated face recognition systems.

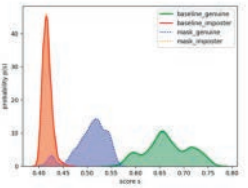
6 Evaluation results

Figure 2 presents the comparison between the baseline (BLR-BLP) genuine and imposter score distributions and the different masked faces experiments (BLR-M1P, BLP-M2P, BLR-M12P) on the three considered face recognition solutions. It is noticeable in all experimental setups that, when comparing masked faces probes to unmasked references, the genuine score distributions strongly shift towards the imposter distributions in comparison to the BLR-BLP setup. This indicates an expected decrease in performance and general trust in the matcher decision, as the separability between genuine and imposter samples decreases. This unwanted shift seems to be slightly stronger when the masked faces are

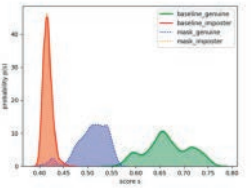
captured under additional artificial illumination (BLR-M2P) when compared to the natural light condition (BLR-M2P). On the other hand, the imposter score distributions do not seem to be significantly affected by the masked probes (BLR-M1P, BLP-M2P, BLR-M12P) in comparison to the unmasked baseline (BLR-BLP).



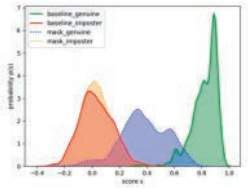
(a) ArcFace: BLR-BLP and BLR-M1P



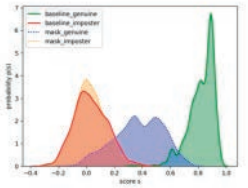
(b) ArcFace: BLR-BLP and BLR-M2P



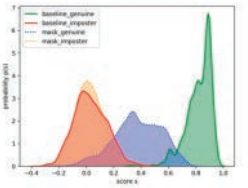
(c) ArcFace: BLR-BLP and BLR-M12P



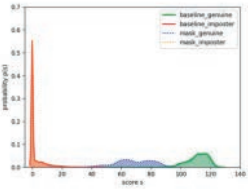
(d) SphereFace: BLR-BLP and BLR-M1P



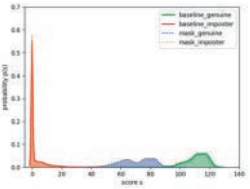
(e) SphereFace: BLR-BLP and BLR-M2P



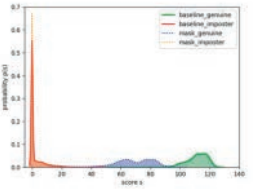
(f) SphereFace: BLR-BLP and BLR-M12P



(g) COTS: BLR-BLP and BLR-M1P



(h) COTS: BLR-BLP and BLR-M2P



(i) COTS: BLR-BLP and BLR-M12P

Fig. 2: The comparison score (similarity) distributions comparing the ”baseline” BLR-BLP genuine and imposter distributions to those of the distributions including ”masked” faces probes (BLR-M1P (a, d, g)), BLR-M2P (b, e, h), BLR-M12P (c, f, i). The shift of the genuine scores towards the imposter distribution is clear when faces are masked for all investigated system (ArcFace(a, b, c), SphereFace (d, e, f), and COTS (g, h, i)).

Tables 2, 3, and 4 present the achieved performance, given by the different evaluation metrics, on all experimental setups by the ArcFace, SphereFace, and COTS solutions, respectively. In all systems, wearing a face mask affected the ability to detect the face properly, resulting in a higher than zero (as in the baseline) FTX. Interestingly, additional illumination (typically from the top) increased the FTX in all systems (BLR-M2P compared to BLR-M1P). This is probably due to the different reflection and shadow patterns induced by the illumination, see samples in Figure 1. The FTX values for the SphereFace and ArcFace in tables 2 and 3 are identical as they both use the MTCNN network for face detection and alignment.

The verification performance (EER, FMR100, FMR1000, ZeroFMR) of the ArcFace and SphereFace is negatively affected when the probe faces are masked (BLR-M1P and BLR-M2P), see tables 2 and 3. This negative effect is stronger when the faces are captured under the effect of artificial illumination (BLR-M2P), probably due to unexpected reflections and shadowing and the fact that the BLR references were captured without such illumination. The reduction in the performance is much more dominant in the SphereFace solution in comparison to the ArcFace. For both systems, the G-mean values decreased significantly when considering the masked probes. This, despite the small size of the evaluation data, indicates a strong negative effect of the masks on the face recognition performance. On the other hand, the I-mean value when considering the masked faces, in comparison to the baseline (BLR-BLP), was not changed under the the ArcFace solution and only slightly changed under the SphereFace solution.

ArcFace	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FTX
BLR-BLP	0.000%	0.000%	0.000%	0.000%	0.666	0.417	0.000%
BLR-M1P	3.163%	3.517%	3.831%	5.069%	0.511	0.417	3.750%
BLR-M2P	5.504%	6.163%	6.628%	7.616%	0.509	0.417	5.833%
BLR-M12P	4.380%	4.888%	5.229%	6.468%	0.510	0.417	4.792%

Tab. 2: The verification performance measures, the G-mean, and I-mean achieved by ArcFace on the different experimental setups. Note the performance degradation induced by the masked face probes.

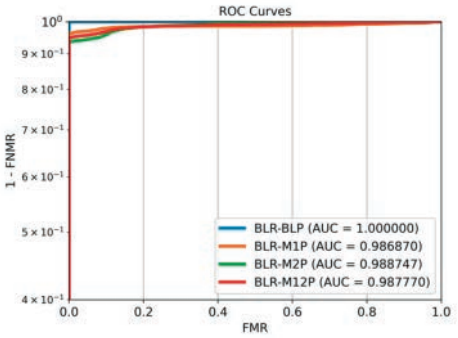
SphereFace	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FTX
BLR-BLP	0.216%	0.065%	0.217%	0.390%	0.825	0.033	0.000%
BLR-M1P	9.312%	27.35%	52.95%	72.91%	0.384	0.026	3.750%
BLR-M2P	12.36%	28.22%	47.66%	73.16%	0.374	0.025	5.833%
BLR-M12P	10.85%	27.86%	50.01%	73.38%	0.380	0.025	4.792%

Tab. 3: The verification performance measures, the G-mean, and I-mean achieved by SphereFace on the different experimental setups. Note the performance degradation induced by the masked face probes.

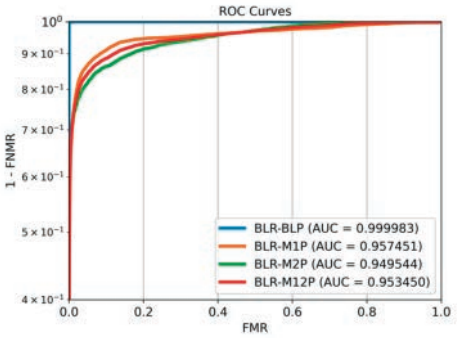
COTS	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FTX
BLR-BLP	0.249%	0.000%	0.251%	0.668%	110.8	2.281	0.000%
BLR-M1P	0.443%	0.304%	0.684%	2.253%	68.09	2.221	2.500%
BLR-M2P	0.004%	0.000%	0.009%	0.050%	70.88	2.298	3.542%
BLR-M12P	0.239%	0.152%	0.341%	1.237%	69.49	2.259	3.021%

Tab. 4: The verification performance measures, the G-mean, and I-mean achieved by COTS on the different experimental setups. Although the change in the performance caused by the masked face probes is insignificant, the shift in the average genuine score towards the imposter scores is very dominant in these cases.

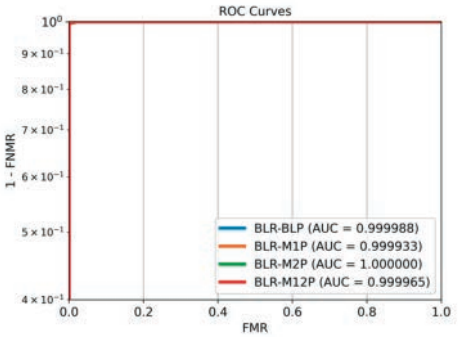
When it comes to verification performance metrics (EER, FMR100, FMR1000, ZeroFMR), the COTS is not significantly affected by masked faces. This is apparent in Table 4, where these performance metrics are not significantly different in all experimental setups. This might be due to the robust and high performance of the COTS solution and the limited size of the evaluation database. However, the change in the G-mean from 110.8 in the BLR-BLP to 69.46 in the BLR-M12P, while maintaining a similar I-mean, indicates a large



(a) ArcFace



(b) SphereFace



(c) COTS

Fig. 3: The verification performance for the three investigated system (ArcFace(a), SphereFace (b), and COTS (c)) is presented as ROC curves. For each of the systems, four curves are plotted to represent the three settings that include ”masked” faces probes (BLR-M1P, BLR-M2P, and BLR-M12P) and the unmasked baseline (BLR-BLP). The area under curve (AUC) is also listed for each of the ROC curves. As in Tables 2, 3, and 4, the effect of masked probes is apparent on the performance of the ArcFace and SphereFace, while the performance of the COTS is almost perfect in all experimental settings (however, with shift in genuine scores values).

change in the separability (between genuine and imposter) in the COTS decisions. This can lead to an increase in the error rate given a larger and more challenging evaluation. Such an evaluation is planned as the data presented in this paper is an initial version of a larger data being collected at the moment. To show the verification performance over a wider range of operation points, Figure 3 presents the ROC curves for the different experimental settings for each of the three investigated systems. Similar conclusions to those established from Tables 2, 3, and 4 can be made. The ArcFace and SphereFace verification performance is effected by the masked probe faces, while the COTS maintains an almost perfect verification performance. However, one must keep in mind the significant shift in the genuine score values in all three systems, as illustrated in Figure 2.

In general, the effect of wearing face masks on the face recognition behaviour is apparent on all investigated systems. The effect is most significant on the genuine scores distribution, rather than the imposter scores distribution. This renders the current face recognition solutions undependable to match masked faces with unmasked faces and, at least, requires re-evaluation.

7 Conclusion

Addressing the wide spread use of face masks as a preventive measure to the COVID-19 pandemic spread, we presented an exploratory study on the effect of wearing masks on face recognition performance in collaborative scenarios. We presented a specifically collected database captured in three different sessions, with and without wearing a mask, and is part of an ongoing effort to gather a larger scale database with realistic variations. We analysed the behaviour of two high-performing academic face recognition solutions and one of the top performing COTS solutions. Our analyses pointed out the significant effect of wearing a mask on comparison scores separability between genuine and imposter comparisons in all the investigated systems. Moreover, we point out a large drop in the verification performance of the academic face recognition solutions, even on a limited evaluation data, when considering masked face probes.

Acknowledgment

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

References

[Da18] Damer, Naser; Wainakh, Yaza; Boller, Viola; von den Berken, Sven; Terhörst, Philipp; Braun, Andreas; Kuijper, Arjan: CrazyFaces: Unassisted Circumvention of Watchlist Face Identification. In: 9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018. IEEE, pp. 1–9, 2018.

[Da19] Damer, Naser; Saladie, Alexandra Mosegui; Zienert, Steffen; Wainakh, Yaza; Terhörst, Philipp; Kirchbuchner, Florian; Kuijper, Arjan: To Detect or not to Detect: The Right Faces to Morph. In: 2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019. IEEE, pp. 1–8, 2019.

[DD16] Damer, Naser; Dimitrov, Kristiyan: Practical View on Face Presentation Attack Detection. In (Wilson, Richard C.; Hancock, Edwin R.; Smith, William A. P., eds): Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016. BMVA Press, 2016.

[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. pp. 4690–4699, 2019.

[Ge17] Ge, Shiming; Li, Jia; Ye, Qiting; Luo, Zhao: Detecting Masked Faces in the Wild with LLE-CNNs. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, pp. 426–434, 2017.

[GP20] Grother Patrick, Ngan Mei, Hanaoka Kayee: Ongoing Face Recognition Vendor Test (FRVT). NIST Interagency Report, 2020.

[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. 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. IEEE Computer Society, pp. 770–778, 2016.

[Hu07] Huang, Gary B.; Ramesh, Manu; Berg, Tamara; Learned-Miller, Erik: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[Li17] Liu, Weiyang; Wen, Yandong; Yu, Zhiding; Li, Ming; Raj, Bhiksha; Song, Le: SphereFace: Deep Hypersphere Embedding for Face Recognition. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6738–6746, 2017.

[Ma06] Mansfield, A: Information technology–Biometric performance testing and reporting–Part 1: Principles and framework. ISO/IEC, pp. 19795–1, 2006.

[Ne] Neurotechnology MegaMatcher 11.2 SDK. "https://www.neurotechnology.com/mm_sdk.html".

[Op16] Opitz, Michael; Waltner, Georg; Poier, Georg; Possegger, Horst; Bischof, Horst: Grid Loss: Detecting Occluded Faces. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III. volume 9907 of Lecture Notes in Computer Science. Springer, pp. 386–402, 2016.

[So19] Song, Lingxue; Gong, Dihong; Li, Zhifeng; Liu, Changsong; Liu, Wei: Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 773–782, 2019.

[Wa20] Wang, Zhongyuan; Wang, Guangcheng; Huang, Baojin; Xiong, Zhangyang; Hong, Qi; Wu, Hao; Yi, Peng; Jiang, Kui; Wang, Nanxi; Pei, Yingjiao; Chen, Heling; Miao, Yu; Huang, Zhibing; Liang, Jinbi: , Masked Face Recognition Dataset and Application, 2020.

[WHM11] Wolf, Lior; Hassner, Tal; Maoz, Itay: Face recognition in unconstrained videos with matched background similarity. In: CVPR 2011. IEEE, pp. 529–534, 2011.

[Yi14] Yi, Dong; Lei, Zhen; Liao, Shengcai; Li, Stan Z.: Learning Face Representation from Scratch. CoRR, abs/1411.7923, 2014.

[Zh16] Zhang, Kaipeng; Zhang, Zhanpeng; Li, Zhifeng; Qiao, Yu: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 11

Soft-Biometrics Estimation In the Era of Facial Masks

Fernando Alonso-Fernandez[1], Kevin Hernandez Diaz[2], Silvia Ramis[3],
Francisco J. Perales[4], Josef Bigun[5]

Abstract: We analyze the use of images from face parts to estimate soft-biometrics indicators. Partial face occlusion is common in unconstrained scenarios, and it has become mainstream during the COVID-19 pandemic due to the use of masks. Here, we apply existing pre-trained CNN architectures, proposed in the context of the ImageNet Large Scale Visual Recognition Challenge, to the tasks of gender, age, and ethnicity estimation. Experiments are done with 12007 images from the Labeled Faces in the Wild (LFW) database. We show that such off-the-shelf features can effectively estimate soft-biometrics indicators using only the ocular region. For completeness, we also evaluate images showing only the mouth region. In overall terms, the network providing the best accuracy only suffers accuracy drops of 2-4% when using the ocular region, in comparison to using the entire face. Our approach is also shown to outperform in several tasks two commercial off-the-shelf systems (COTS) that employ the whole face, even if we only use the eye or mouth regions.

Keywords: Soft-Biometrics, Periocular, Gender, Age, Ethnicity.

1 Introduction

Recent research has explored the use of ancillary information, known as soft biometrics, which includes attributes like gender, age, ethnicity, etc. [DER16]. While they may not be sufficiently distinctive to allow accurate recognition, they can be used in a fusion framework to complement the primary system [Go18]. Automated soft-biometrics extraction has other applications as well, such as reducing the search space of subjects in large databases, locating specific individuals based on such semantic attributes, providing age-dependant access control, or customizing advertisements or customer recommendations [DER16].

Face is a natural way to recognize many soft-biometrics indicators. However, in unconstrained conditions, it may be partially occluded, accidentally or intentionally, as for example by the use of masks. Accordingly, we address the challenge of estimating soft-biometrics indicators when only images of face parts are available. This has been suggested in several studies with traditional features such as Local Binary Patterns or Histograms of Oriented Gradients [AFB16]. Here, we leverage the power of Convolutional Neural Networks (CNNs) pre-trained in the context of the ImageNet challenge with more than a million images to classify images into 1000 object categories. Based on [Ng18], the

[1] School of Information Technology, Halmstad University, Sweden, feralo@hh.se
[2] School of Information Technology, Halmstad University, Sweden, kevin.hernandez-diaz@hh.se
[3] Computer Graphics and Vision and AI Group, University of Balearic Islands, Spain, silvia.ramis@uib.es
[4] Computer Graphics and Vision and AI Group, University of Balearic Islands, Spain, paco.perales@uib.es
[5] School of Information Technology, Halmstad University, Sweden, josef.bigun@hh.se

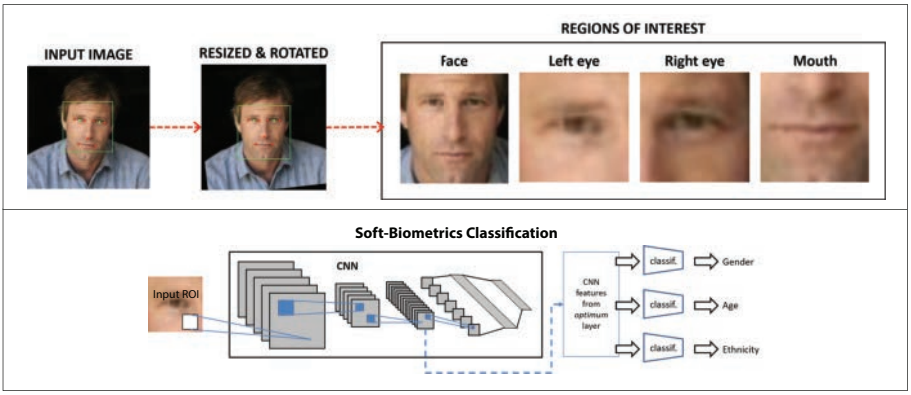


Fig. 1: Top: Extraction of the regions of interest. Bottom: Soft-biometrics classification framework.

authors in [HDAFB18, Al19] investigated the use of these off-the-shelf CNNs for periocular recognition, eliminating the necessity of designing and training new networks. Here, we further investigate their behaviour in soft-biometrics classification. Our experiments show that these off-the-shelf features are capable of measuring soft-biometrics using only the ocular or mouth regions, with negligible accuracy drops or even better performance in comparison to using the whole face. The proposed approach also compares favourably with two commercial off-the-shelf systems (COTS), outperforming them in several tasks.

2 Soft-Biometrics Classification Approach

We extract features from different regions (Figure 1): face, left/right periocular, or mouth. For feature extraction, the following networks are used: AlexNet [KSH12], ResNet50 and ResNet101 [He16], DenseNet201 [Hu17], VGG-Face [PVZ15], and MobileNetv2 [Sa18]. These networks have gained in sophistication and depth, starting from AlexNet (with only 5 convolutional layers), to ResNet (50/101 layers) and DenseNet (201 layers). The latter were made possible thanks to concepts like residual connections [He16] and densely connected architectures [Hu17], with allowed the training of deeper networks. We also employ VGG-Face. Based on the generic VGG16, it is trained to recognize faces, so we believe that it can provide effective recognition in our tasks with data from facial regions. Finally, we use the network MobileNetv2, designed to have a smaller size while keeping accuracy. With these choices, we aim at comparing networks of different depths, and a network trained with faces as well. In using them, images are fed into each CNN. But instead of using the vector from the last layer, we employ as descriptor the intermediate layer identified as giving the best performance in periocular recognition [HDAFB18, Al19]. Since we will employ a similar type of data, we speculate that these layers will be useful for soft-biometrics as well. In particular, we use the layers: 14 (AlexNet), 73 (ResNet50), 165 (ResNet101), 223 (DenseNet201), 25 (VGG-Face) and 121 (MobileNetv2). Classification with each network is then done by training a linear Support Vector Machine (SVM) with the extracted feature vectors [Va95]. The complete procedure is shown in Figure 1 (bottom), whereas Table 1 indicates the size of the feature vector for each network.

Network	Layer	Size
AlexNet	14	43264
ResNet50	73	100352
ResNet101	165	50176
DenseNet201	223	6272
VGG-Face	25	100352

Network	Layer	Region	Size
MobileNetv2	121	-	7840
MobileNetv2	121	face	4763
+ PCA		left eye	4332
		right eye	4327
		mouth	4396

Tab. 1: Size of the feature vector per classification network.

Attribute					
Gender	Male (77.6%)	Female (22.4%)			
Age	Baby (<1%)	Child (<1%)	Youth (12.9%)	Adult (62.9%)	Senior (23.6%)
Ethnicity	White (81.6%)	Black (3.8%)	Asian (5.5%)	Indian (2.4%)	Other (6.7%)

Tab. 2: Statistics of soft-biometrics attributes of the LFW database.

3 Database and Protocol

We use the Labeled Faces in the Wild (LFW) database [Hu07]. It contains images of celebrities from the web with a large range of variations in pose, lightning, expression, etc. In particular, we use 12007 images, for which annotation of face landmarks is available. All images are rotated w.r.t. the axis crossing the eyes, and resized to an eye-to-eye distance of 42 pixels (average of the database). Then, a face image of 109×109 is extracted, together with the two periocular regions (43×43 each), and the mouth (49×49). Images are further resized to the input size of the networks. An example of this procedure is given in Figure 1. To train and evaluate our classification approach, we employ the ground-truth of [Go18]. Table 2 indicates the attributes employed and the statistics of the database. When there are more than two classes, a one-vs-one multi-class approach is used. For every feature and N classes, $N(N-1)/2$ binary SVMs are used. Classification is made based on which class has most number of binary classifications towards it (voting scheme). Evaluation is done with k-fold cross-validation (k=5), with k sets containing the same number of (non-overlapped) people. On each iteration, a set is retained for validation, and the remaining $k-1$ sets are used to train the SVMs. The average accuracy of the k iterations are then reported. The software employed was Matlab r2019a, which contains pre-trained models of all the CNNS, except VGG-Face which is from the Caffe Model Zoo.

4 Results

The performance of our soft-biometrics classification approach is reported in Figures 2-4 for gender, age, and ethnicity respectively. Accuracy is reported for each class (images of the class classified correctly), and for the whole database (images of the database classified correctly). We provide results using as input: *i*) the whole face, *ii*) the left/right eye separately, *iii*) both eyes together (by concatenating feature vectors), and *iv*) the mouth region.

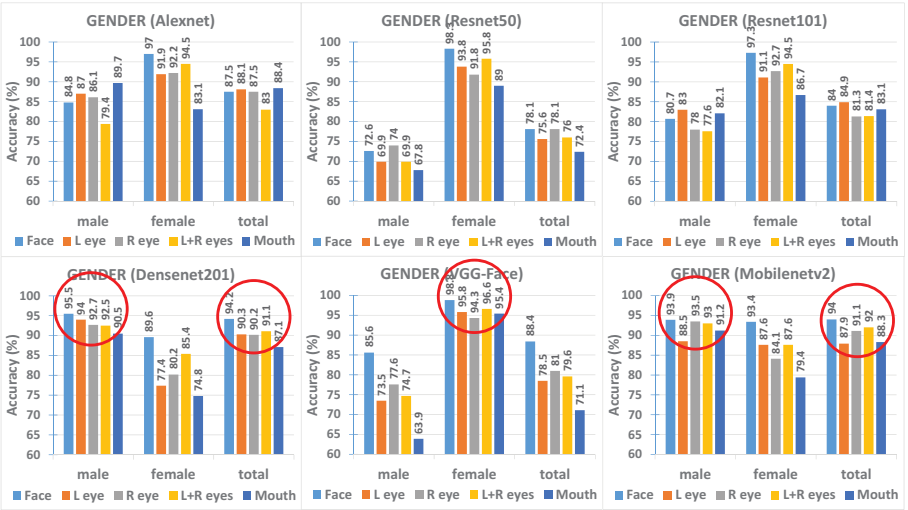


Fig. 2: Accuracy of gender estimation using different facial regions.

The size of some age groups ("baby", "child") is very small, see Table 2, so these groups been merged with the class "youth" into a single class that we call "minors".

The red circles indicate the top results for each class. A quick look reveals that three networks concentrate the top results (with few exceptions): DenseNet201, VGG-Face and MobileNetv2. The best networks overall ('total' accuracy) are DenseNet201 and MobileNetv2. This is interesting, since DenseNet201 is the deepest network employed, while MobileNetv2 is a lighter network designed to have much less depth and parameters. With DenseNet201, gender is estimated with an accuracy of 87.1-94.2% (depending on the image region), while age is estimated with 57.6-62%, and ethnicity with 76.8-81.6%. With MobileNetv2, gender is estimated with an accuracy of 87.9-94%, age with 55.5-63.8%, and ethnicity with 70.3-80.5%. It is also relevant that VGG-face does not systematically outperform the other networks, even if it is trained with facial data. DenseNet201 and MobileNetv2 are also the best network with the classes having more samples (Table 2): gender-male, age-adult, and ethnicity-white classes. On the other hand, VGG-Face wins with the classes that are less represented; a downside though is that its performance with the biggest classes is poor. The latter is also seen in the ResNet variants.

Interestingly, the feature vectors of DenseNet201 and MobileNetv2 are the smallest among those employed (Table 1). Therefore, a bigger feature vector does not correlate with a better performance, but the opposite. Also, MobileNetv2 stands out as a very balanced network, with top results with the biggest classes, and also relatively good performance with the others (with very few exceptions like ethnicity-indian or ethnicity-other classes, whose performance is very poor with any network). Given that the networks employed have not been specifically trained for soft-biometrics, and to eliminate feature redundancy, we carry out dimensionality reduction by Principal Component Analysis (PCA) [Jo02]. We retain the elements with 99% of the variance, with the PCA basis learnt using images from

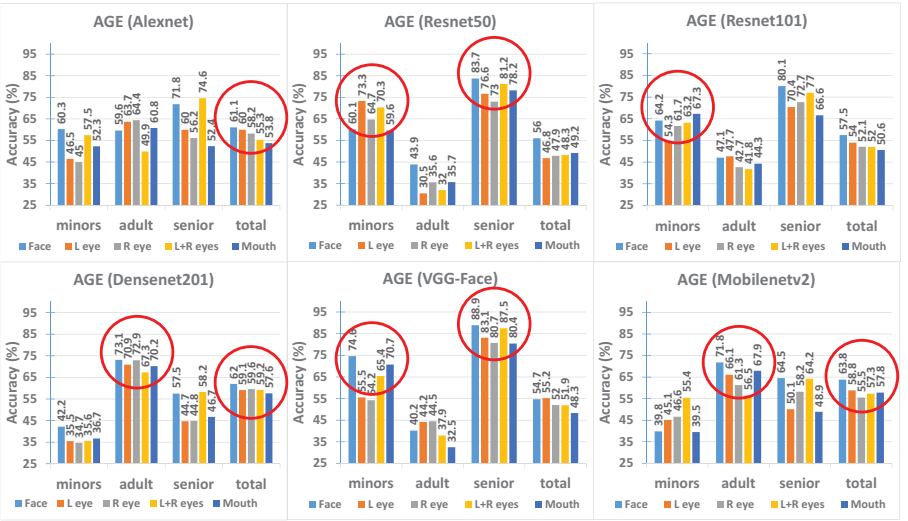


Fig. 3: Accuracy of age estimation using different facial regions.

the training set on each validation iteration. In our experiments, we have observed that PCA provides further performance improvement with DenseNet201 and MobileNetv2 in the majority of classes. On the other hand, results with the other networks are not consistent, showing improvement with some classes, while decreasing substantially in others. Due to space, we only show results of MobileNetv2 (Table 3). Also, Table 1 (right) gives the average number of retained coefficients for the different regions.

As it can be observed in Table 3, in overall terms ('total' columns), PCA provides an extra improvement. The performance of the biggest classes (gender-male, age-adult, and ethnicity-white) is better, and improvements happen as well with several less-represented classes. It happens though that some small classes worsen after PCA, e.g. age-senior, ethnicity-black, or ethnicity-other. Regarding the use of different facial regions, it can be observed that using only the periocular or mouth regions is not necessarily worse than using the whole face. This is not only seen with MobileNetv2 (Table 3), but with other networks as well (Figures 2-4). When estimating gender with MobileNetv2, the best accuracy is obtained with the whole face (95.8%). With a combination of both eyes, accuracy is just 2.4% below (93.4%), and with only one eye, it drops a further 0.8% only (92.6%). Accuracy with only the mouth region is also comparably good (90.5%), although its accuracy with the gender-female class is much worse than the other facial regions. In a similar vein, the whole face provides the best overall performance in age (64.5%) and ethnicity (83.3%) estimation, and the use of facial parts results in a small accuracy drop only. Age with only the mouth is estimated with an accuracy of 59.6%, which goes up to 60% when both eyes are used, and even better with the left eye only (60.2%). Similarly, ethnicity with both eyes or the mouth is estimated with an accuracy of 81.3%/81.5%, and even better with the right eye only (82.9%). It is worth noting as well that combining both eyes does not necessarily

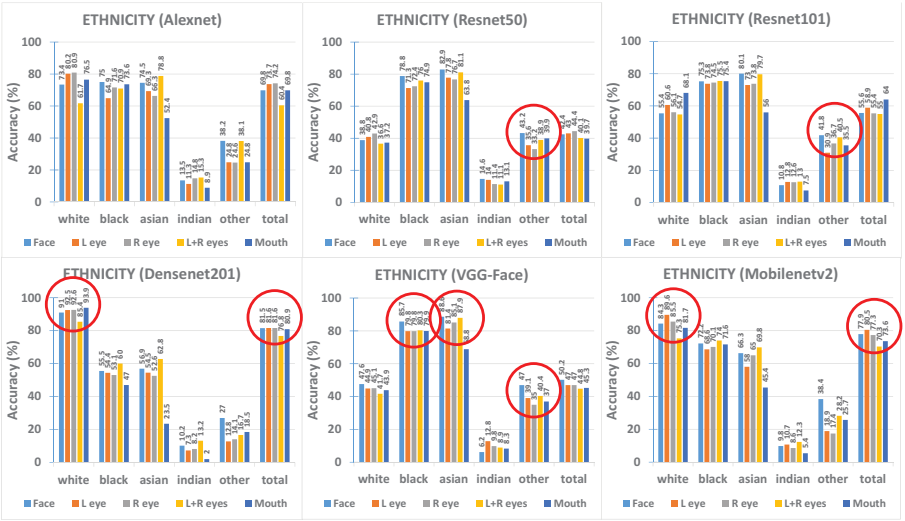


Fig. 4: Accuracy of ethnicity estimation using different facial regions.

produces better accuracy, in comparison to using one eye only. In addition, uncorrelated areas such as the eye or mouth provides a relatively similar performance.

We also provide (Table 4) the results of two COTS systems, Face++[3] and Microsoft Cognitive Toolkit[4], given in [Go18]. These systems estimate soft-biometrics attributes based on deep learning architectures. The results in Table 4 have been obtained using images of the whole face. Note that not all the classes employed in this paper are provided. Ethnicity is only given by Face++, giving only the classes white (caucasian), black and asian. Regarding age, the results in [Go18] are separated by the five age groups of Table 2. By comparing Tables 3 and 4, we observe that the performance of our suggested framework using MobileNetv2 outperforms the gender estimation of these COTS systems. Regarding age estimation, the COTS systems are better for age classes involving minors (which represent only about 13% of the data), but they show poorer performance with age-adult or age-senior groups. Regarding ethnicity, our approach outperforms the COTS systems for white and black classes. It is also worth noting that in the classes where our approach outperforms the COTS systems, the superiority is observed as well if we only employ the eye or mouth regions.

5 Conclusions

We suggests the use of off-the-shelf CNN architectures, pre-trained in the context of the ImageNet Large Scale Visual Recognition Challenge, for the task of soft-biometrics classification with facial images. More importantly, giving the current context where face en-

[3] https://www.faceplusplus.com/
[4] https://www.microsoft.com/cognitive-services/

	GENDER			AGE			
	male	female	total	minors	adult	senior	total
face	93.9	**93.4**	94	39.8	71.8	**64.5**	63.8
face + PCA	**97.6**	90.1	**95.8**	45	**75.6**	53.1	**64.5**
left eye	88.5	87.6	87.9	45.1	66.1	50.1	58.8
left eye + PCA	95.8	80.8	92.5	46.1	69.4	46.7	60.2
right eye	93.5	84.1	91.1	46.6	61.3	58.2	55.5
right eye + PCA	94.6	85.3	92.6	39.9	70.5	46	57.8
both eyes	93	87.6	92	**55.4**	56.5	64.2	57.3
both eyes + PCA	94.6	89.7	93.4	45.9	72.8	45.9	60
mouth	91.2	79.4	88.3	39.5	67.9	48.9	57.8
mouth + PCA	95.2	74.6	90.5	41.9	71.9	44	59.6

	ETHNICITY					
	white	black	asian	indian	other	total
face	84.3	72.2	66.3	9.8	**38.4**	77.9
face + PCA	91.1	**78.1**	66.8	7.7	32	**83.3**
left eye	89.6	68.6	58	10.7	18.9	80.5
left eye + PCA	90.2	68.5	60.4	**13.4**	23.5	81.4
right eye	85.5	70.1	65	8.6	17.4	77.3
right eye + PCA	**93.2**	63.8	55.3	8.2	17	82.9
both eyes	75.3	74	69.8	12.3	28.2	70.3
both eyes + PCA	88.8	70.8	**76.8**	11.1	25.3	81.3
mouth	81.7	71.6	45.4	5.4	25.7	73.6
mouth + PCA	92	69.5	38.8	4.3	21.7	81.5

Tab. 3: MobileNetv2 network: Accuracy of soft-biometrics estimation with and without PCA reduction using different facial regions. For each region, the best accuracy (between using/not using PCA) is highlighted with a grey background. The best overall accuracy of each class is marked in bold.

gines are forced to work with images of people wearing masks, we evaluate the feasibility of using partial images containing only the ocular or mouth regions (Figure 1). In this paper, we test popular generic architectures, with features extracted from intermediate layers identified in previous studies as providing good person recognition with ocular images. Prediction is then done with SVM classifiers. They are evaluated with 12007 annotated images of the LFW database [Hu07, Go18]. Our results indicate the possibility of performing soft-biometrics classification using images containing only the ocular or mouth regions, without a significant drop in performance in comparison to using the entire face. An overall accuracy of 95.8/64.5/83% in gender/age/ethnicity estimation is obtained with images of the entire face using the MobileNetv2 architecture. Using only images of one eye, the best accuracy in these tasks is 92.6/60.2/82.9% respectively, and using images of the mouth area, we obtain an accuracy of 90.5/59.6/81.5%. The proposed approach also compares well with two COTS systems by Face++ and Microsoft, outperforming them in the gender estimation task, and in several classes of the age and ethnicity tasks.

A limitation to overcome is the class imbalance of our database. Also, the CNN layers employed were optimized for periocular recognition, but it might be that the best layer for soft-biometrics or for the entire face or the mouth region is different. We are also looking

GENDER

Face++			Microsoft		
male	female	total	male	female	total
92.2	87.5	91.1	93.5	91.1	92.9

AGE

Face++					
baby	child	youth	adult	senior	total
100	53.2	81.4	32	33.4	38.8

ETHNICITY

Face++					
white	black	asian	indian	other	total
88.3	76.2	83.1	-	-	87.4

AGE

Microsoft					
baby	child	youth	adult	senior	total
100	45.2	92.2	52.5	59.6	59.3

Tab. 4: Performance of Face++ and Microsoft COTS [Go18].

into fine-tuning CNN architectures to do the classification directly, thanks to newer annotated repositories [MFV19]. We also foresee that improvements can be obtained by joint estimation of soft-biometrics indicators by sharing weights between different networks, since a single facial feature carry information about different soft-biometrics [DER16].

Acknowledgment

This work was partly done while F. Alonso-Fernandez. was a visiting researcher at the University of Balearic Islands (UIB), funded by the visiting lecturers program of the UIB. Authors F. Alonso-Fernandez, K. Hernandez-Diaz and J. Bigun would like to thank the Swedish Research Council for funding their research. Authors F. J. Perales and S. Ramis would like to thank the project PERGAMEX RTI2018-096986-B-C31 (MINECO/AEI/ERDF, EU) and the project PID2019-104829RA-I00 / AEI / 10.13039/501100011033 (MICINN). Part of the computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at NSC Linköping. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[AFB16] Alonso-Fernandez, F.; Bigun, J.: A survey on periocular biometrics research. Pattern Recognition Letters, 82:92–105, 2016.

[Al19] Alonso-Fernandez, F.; Raja, K. B.; Raghavendra, R:; Busch, C.; Bigün, J.; Vera-Rodríguez, R.; Fiérrez, J.: Cross-Sensor Periocular Biometrics: A Comparative Benchmark including Smartphone Authentication. CoRR, abs/1902.08123, 2019.

[DER16] Dantcheva, A.; Elia, P.; Ross, A.: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE TIFS, 11(3):441–467, 2016.

[Go18] Gonzalez-Sosa, E.; Fierrez, J.; Vera-Rodriguez, R.; Alonso-Fernandez, F.: Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation. IEEE TIFS, 13(8):2001–2014, August 2018.

[HDAFB18] Hernandez-Diaz, K.; Alonso-Fernandez, F.; Bigun, J.: Periocular Recognition Using CNN Features Off-the-Shelf. In: Proc BIOSIG. pp. 1–5, Sep. 2018.

[He16] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep Residual Learning for Image Recognition. In: Proc CVPR. pp. 770–778, June 2016.

[Hu07] Huang, G. B. et al.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. TR 07-49, Univ of Massachusetts, Oct 2007.

[Hu17] Huang, G. et al.: Densely Connected Convolutional Networks. Proc CVPR, 2017.

[Jo02] Jolliffe, Ian: Principal component analysis. Springer Verlag, New York, 2002.

[KSH12] Krizhevsky, A. et al.: ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Informat Proc Systems 25. Curran Associates, Inc., 2012.

[MFV19] Morales, A.; Fierrez, J.; Vera-Rodríguez, R.: SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. CoRR, abs/1902.00334, 2019.

[Ng18] Nguyen, K.; Fookes, C.; Ross, A.; Sridharan, S.: Iris Recognition With Off-the-Shelf CNN Features: A Deep Learning Perspective. IEEE Access, 6:18848–18855, 2018.

[PVZ15] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.: Deep Face Recognition. Proc BMVC, 2015.

[Sa18] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proc CVPR. pp. 4510–4520, 2018.

[Va95] Vapnik, V. N.: The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 21

Watchlist Adaptation: Protecting the Innocent

Manuel Günther[1], Akshay Raj Dhamija[2], Terrance E. Boult[2]

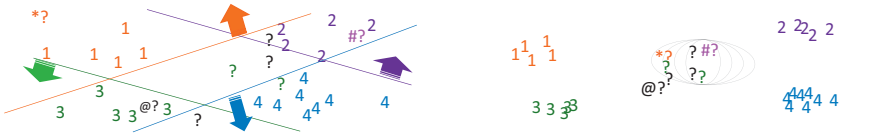
Abstract: One of the most important government applications of face recognition is the watchlist problem, where the goal is to identify a few people enlisted on a watchlist while ignoring the majority of innocent passersby. Since watchlists dynamically change and training times can be expensive, the deployed approaches use pre-trained deep networks only to provide deep features for face comparison. Since these networks never specifically trained on the operational setting or faces from the watchlist, the system will often confuse them with the faces of innocent non-watchlist subjects leading to difficult situations, e.g., being detained at the airport to resolve their identity. We develop a novel approach to take an existing pre-trained face network and use adaptation layers trained with our recently developed Objectosphere loss to provide an open-set recognition system that is rapidly adapted to the gallery while also ignoring non-watchlist faces as well as any background detections from the face detector. While our adapter network can be quickly trained without the need of re-training the entire representation network, it can also significantly improve the performance of any state-of-the-art face recognition network like VGG2. We experiment with the largest open-set face recognition dataset, the UnConstrained College Students (UCCS). It contains real surveillance camera stills including both known and unknown subjects, as well as many non-face regions from the face detector. We show that the Objectosphere approach is able to reduce the feature magnitude of unknown subjects as well as background detections, so that we can apply a specifically designed similarity function on the deep features of the Objectosphere network, which works much better than the direct prediction of the very same network. Additionally, our approach outperforms the VGG2 baseline by a large margin by rejecting the non-face data, and also outperforms prior state-of-the-art open-set recognition algorithms on the VGG2 baseline data.

Keywords: Open-Set Face Recognition, Watchlist, Gallery Adaptation.

1 Introduction

In recent years, face biometric systems using deep networks have matured into an age of high performance. These advances have lead face biometrics into daily-use applications such as access control for mobile devices or tagging friends on social media, but also an increasing usage by governments and law-enforcement for security can be observed. However, there remains at least one application for which their performance is insufficient and where errors impact innocent citizens: watchlists. A watchlist is an open-set problem and, because most people are not in the gallery of subjects of interest, the system must operate at a very low false alarm rate to reject the predominately unknown people. Recently, one of the vendors faced considerable criticism for matching US congress members to mugshots of criminals [Ro17]. That research was an eye-opener on the state of such commercial recognition systems since false alarms can substantially bias the interaction of security personnel with the person in question while increasing the responsibility for officers to verify the outputs of the system. Consequently, the latest NIST evaluations [GNH19] also include watchlist protocols, though only on images in controlled conditions.

[1] University of Zurich, Department of Informatics, Binzmühlestrasse 14, CH-8050 Zurich, guenther@ifi.uzh.ch
[2] University of Coloado Colorado Springs, Vision and Security Technology Lab, 1420 Austin Bluffs Parkway, CO-80933 Colorado Springs, {adhamija,tboult}@vast.uccs.edu



(a) Linear classifiers have unbounded open-space risk and unknowns are often misclassified

(b) Watchlist adaption refines features into tighter clusters and brings unknowns to the center

Fig. 1: OBJECTOSPHERE. *(a) Features from pre-trained face recognition networks often leave innocent subjects (?) near watchlist subjects (numbers), so any distance-based rejection fails. Gallery adaptation, e.g. using linear SVMs, can leave innocent subjects (*?, ?? #?) associated with a watchlist subject, often with high confidence. In our watchlist adaptation approach (b), we use an Objectosphere-based deep feature adapter to learn to specialize the features to the watchlist samples while mapping innocent unknown (*?, @? #?) samples to be near the origin. Watchlist subjects (numbers) have more compact representations, and when an innocent unknown (?) is presented, it has reduced magnitude and its weighted cosine distance to any class allows for rejection. Known samples will still map to be near their training samples, and the weighted cosine will correctly match them.*

In our recent work [DGB18], we showed that one of the difficulties with open-set recognition in deep networks is that deep features for unknown inputs will often map near or directly on top of the features of known inputs, as indicated in Fig. 1. If the features overlap, no distance-based algorithm can separate them. Thus we argue that for dealing with unknown subjects in a watchlist, we ideally want to learn deep features that separate the known subjects from unknown inputs. Note that unlike prior work, our "unknowns" include very different objects, e.g., backgrounds that make it through the face detector as well as faces of unknown subjects. The latter are very similar to the known subjects, which makes this a more difficult separation.

Currently, face research/practice eschews "training" on the gallery [Lu12]. That deeply ingrained decision is more folklore than science and we argue that difficult open-set problems such as watchlists require adapting the deep features to the specific gallery, what we call *watchlist adaptation*. There are two primary arguments against gallery training: cost and generalizability. For large face recognition, such as passport or visa management, there are many people in the gallery, and the gallery is constantly changing so training on it is impractical. However, watchlists tend to be small and rarely modified, and as we shall show in this paper, with the right design, retraining/fine-tuning for a modified watchlist is quick.

The question of generalizability is more subtle. Systems are trained across demographics to ensure that they generalize well. To obtain such a large training dataset, the data is not likely consistent with the domain of the watchlist problem. We argue that face watchlist adaptation is not generalization but a rather proper specialization to the operational domain.

Our contributions in this paper are: *(a)* We develop the first system with watchlist adaptation, providing features tuned to separate the watchlist identities from unknown identities and objects. *(b)* We develop the first novel approach to adapt a pre-trained network using Objectosphere loss. *(c)* We demonstrate that the Objectosphere adapter learns feature representations that are more robust and help to protect innocents in watchlist scenarios. *(d)* We provide state-of-the-art results on the UCCS face watchlist dataset.

2 Related Work

Since deep learning was introduced to face recognition, all modern face recognition algorithms [PVZ15, SKP15, CPC16, Sa16, Ca18, De19] rely on deep neural networks (DNNs). Many algorithms implemented special ways of training the networks in order to provide better performance on difficult images, such as triplet loss embedding [PVZ15, Sa16] or different network topologies [SKP15, Ca18]. Since these networks require large amounts of training data, usually these algorithms are trained on secondary datasets [Ba17, Ca18] that cannot have overlapping identities with tested datasets. Deep feature representations extracted from the penultimate layer are compared using simple distances such as Euclidean [PVZ15] or cosine [Sa16, CPC16, De19]. While these networks provide brilliant performance on imagery with relatively high quality, they cannot handle facial images with low (optical) resolution, or even background detections of the face detector. Generally, images with difficult content have very similar deep features and cluster in the center of the deep feature space [O'18]. Thus, none of these networks is able to reject background detections in a principled way and, therefore, they cannot be applied in real-world applications where (false) alarms need to be handled by a human operator.

A few prior works also saw the need for gallery-related training. Klare *et al.* [Kl15] argue that "training could occur on an active gallery to learn the nuances of subjects that are labeled in a gallery" but did not experiment with such gallery/subject-specific modeling. Chowdhury *et al.* [Ch16] trained one-vs-rest linear SVMs on the gallery. In neither case did they explore gallery specific features or design to separate the features of known and unknown inputs.

The most well-known open-set face recognition datasets come from the IARPA Janus benchmark (IJB) series [Kl15, Wh17, Ma18]. The biggest issue with the IJB datasets is that their protocols only include detected and manually marked faces but no background detections. In contrast, the UnControlled College Students (UCCS) dataset [SB13] and its corresponding protocol [Gü17b] mandate for faces to be detected as part of the recognition pipeline. Hence, the recognition system needs to classify both background detections and unknown faces as unknown. Due to this unique property and its true open-set nature, we use this dataset for our experiments.

3 Approach

Watchlist is a typical open-set recognition scenario where a probe may include an unknown identity. The system should only provide an alert if the probe belongs to one of the known subjects from the gallery G, but not when the probe sample is of an unknown subject $u \in U$ with $G \cap U = \emptyset$. Given a probe sample x_p of subject $g \in G$, the system D needs to produce $D(x_p) \rightarrow g$. If x_p does not belong to any subject in G, the system needs to produce $D(x_p) \rightarrow U$, even when the system never saw this specific subject.

A real-world face watchlist system consists of two sub-systems, i.e., $D = D_d \rightarrow D_r$ where D_d is a detection system and D_r is a representation system that is used to represent the output of D_d for recognition. Thus, the performance of the complete system D is tied to the performance of the detector D_d, and the representation system D_r should act as the last line of defense for overcoming the drawbacks of D_d. Therefore, D_r is susceptible to two types

of samples that it should identify as unknown, i.e., when x_p is a face that does not belong to G, or when x_p is not even a face. In either case, D_r should be able to mark this probe image as not belonging to one of the known faces.

Traditionally, the use of deep networks in face watchlists is limited to representational networks, which enable researchers to decouple the training and the testing pipelines and recognize subjects that the network was not initially trained to identify. During enrollment, representations $R_g = D_r(x_g)$ are obtained for faces belonging to subject $g \in G$. These representations are used to create a gallery template G_g for the subject g. To avoid enrolling bad samples, the detection step D_d is either avoided by hand-labeling the face or at least monitored. During inference, a representation $R_p = D_d \rightarrow D_r(x_p)$ is obtained for a probe image x_p and a similarity score $s(G_g, R_p)$ is calculated between the representations of the probe and the gallery. This score is then thresholded in order to reject probe samples with low similarity to all gallery templates as unknown.

In this paper, we present a new approach to the watchlist problem, where we use the gallery for training new features so that we separate feature representations for persons of interest from representations of unknown samples. Since this approach creates a drastic difference between gallery subjects and unknown faces, it is not possible to perform enrollment for a new subject without retraining. Fortunately, since we rely on features extracted from another representational network, retraining the network is fast and could be performed whenever a new subject needs to be enrolled.

3.1 Training

We use a secondary network (D_c) containing multiple fully connected layers to classify a given feature representation $R = D_r(x)$. We use two different loss functions to train D_c, namely the Objectosphere loss as introduced in [DGB18] and the standard softmax loss with an additional background class, which is often seen for training object detectors.

Objectosphere The Objectosphere loss introduced in [DGB18] is based on the entropic open-set loss J_E. The entropic open-set loss works similarly to the traditional softmax loss for the samples x_g belonging to the known subjects, where each node S_g of the softmax output represents one of the G known subjects. Unknown samples x_u are considered as equal members of each of the possible classes:

$$J_E(x) = \begin{cases} -\log S_g(D_c(R)) & \text{if } x \text{ belongs to } g \\ -\frac{1}{G}\sum_{g'=1}^{G} \log S_{g'}(D_c(R)) & \text{if } x \text{ is unknown} \end{cases} \tag{1}$$

It addition to J_E, Objectosphere applies a constraint on the magnitudes of the features representations. For unknown samples, the objective function forces the magnitudes of the penultimate layer $D_c(R)$ to be close to zero, while pushing the feature magnitudes of known samples to at least ξ, a predefined hyperparameter, which we have set to $\xi = 5$ in our experiments:

$$J_R = J_E + \lambda \begin{cases} \max(\xi - ||D_c(R)||, 0)^2 & \text{if } x \text{ is known} \\ ||D_c(R)||^2 & \text{if } x \text{ is unknown} \end{cases} \tag{2}$$

Softmax For comparison, we apply a technique that is commonly used in object detectors. Similar to the above, we utilize softmax to classify a given input as one of the subjects present in the gallery, and we add an additional output node for the unknown samples. Hence, the last layer of the softmax network has $|G|+1$ nodes.

3.2 Inference

During inference, a representation for the probe image x_p is obtained using the representational network, i.e., $R_p = D_r(x_p)$. This representation is then fed into the secondary network D_c to identify the sample as belonging to one of the subjects in the gallery or not belonging to any of them. We achieve this using the following two approaches:

Classification In this approach, we use the scores of the softmax layer S_g of the classification network D_c to link the probe to a known subject: $s(g,x_p) = S_g(D_c(R_p))$. To obtain an open-set measure, we threshold the softmax score at θ and reject the probe sample as unknown if the softmax score is below θ.

Similarity This approach is the traditional use of deep networks as feature extractors for face recognition. Since our classification network D_c contains multiple fully connected layers, it is able to learn its own representation of the incoming samples. As common, we remove the last fully-connected and the softmax layers of the network and extract $P_p = D_c(R_p)$ from the deep feature layer of the network. We enroll gallery templates G_g by averaging the normalized gallery features $P_g = D_c(R_g)$ of each known subject. For inference, we compute similarity scores between the probe and the gallery using two different similarity functions. First, we compute the cosine similarity $\cos(G_g, P_p)$ between gallery template G_g and probe feature P_p. Since Objectosphere specifically aims at manipulating the magnitude of the deep features P_p, we also multiply the cosine similarity with the magnitude of the probe feature:

$$\mathrm{mcos}(G_g, P_p) = \cos(G_g, P_p) \cdot ||P_p|| = \frac{{G_g}^T P_p}{||G_g||} \tag{3}$$

As before, the maximum similarity to any gallery template is thresholded to reject probe samples as unknown.

4 Experiments

4.1 Evaluation

To evaluate the open-set face recognition performance, we employ an adaptation of the detection and identification rate (DIR) curve [Gü17a], which usually is plotted against the probability of false alarms [PGM11]. The DIR (here we only evaluate the DIR at rank 1) is computed solely on the probe samples of known subjects $\mathbf{K}$. In the DIR, we consider probes to be correctly identified if the similarity to the correct subject g^* is the highest and above similarity threshold θ:

$$\begin{aligned}\mathrm{DIR}(\theta) = \frac{1}{|\mathbf{K}|} \Big| \Big\{ P_p \mid \arg\max_g \; s(G_g, P_p) = g^* \wedge \\ s(G_{g^*}, P_p) \geq \theta;\; P_p \in \mathbf{K} \Big\} \Big|.\end{aligned} \tag{4}$$

(a) Sunny Day (b) Snowy Day

Fig. 2: UCCS DATASET EXAMPLES. *Two images show examples from the UCCS dataset [Gü17b] including their ground-truth bounding boxes and labels. Identical subjects are marked with identical color, while unknown subjects are marked in white.*

The original definition of the DIR curve plots the detection and identification rate (4) over the probability of false alarms $\mathscr{P}_{\text{FA}}$ [PGM11]. The $\mathscr{P}_{\text{FA}}$ is computed on the unknown samples $\mathbf{U}$ only. A false alarm is issued when the similarity of an unknown probe sample P_p to any of the known subjects G_g is larger than θ:

$$\mathscr{P}_{\text{FA}}(\theta) = \frac{1}{|\mathbf{U}|}\left|\left\{P_p \mid \max_g s(G_g, P_p) \geq \theta;\ P_p \in \mathbf{U}\right\}\right|. \tag{5}$$

Using the $\mathscr{P}_{\text{FA}}$ in our evaluation has the issue that the number of unknown samples $\mathbf{U}$ might vary based on the quality of the employed face detector. Hence, a poor face detector that provides many background detections that are easy to reject by the face recognition system would be favored since it lowers the $\mathscr{P}_{\text{FA}}$. Therefore, in [Gü17b] we only computed the total number of false alarms (which we called false identifications), i.e., without normalizing by the number of unknown samples $|\mathbf{U}|$. Unfortunately, the total number of false alarms is not very intuitive and might vary based on the number of probe images. Hence, we divide the number of false alarms by the total number of probe *images* $\mathbf{I}$, where each image contains several probe faces, to obtain the average number of false alarms per image (FAI):

$$\text{FAI}(\theta) = \frac{1}{|\mathbf{I}|}\left|\left\{P_p \mid \max_g s(G_g, P_p) \geq \theta;\ P_p \in \mathbf{U}\right\}\right|. \tag{6}$$

We believe that this metric is best suited for selecting a threshold θ according to specific requirements. For example, if a CCTV camera captures an image every 6 seconds, and we want to limit the impact on innocent subjects by needing human operator intervention once every ten minutes, then the threshold θ should be based on an FAI of 0.01.

4.2 Dataset and Experimental Setup

We evaluate our experiments on the validation set of the UCCS dataset [Gü17b], examples of which are shown in Fig. 2. We use the source code package[3] of the challenge, which includes the evaluation scripts that we used in our evaluation. We actually found a small bug in the evaluation code, which we corrected. Additionally, we modified the false alarms axis to divide by the number of probe images to arrive at the FAI.

[3] http://pypi.org/project/challenge.uccs

In our experiments, we use the publicly available MTCNN2 face detector [Zh16] and the VGG2 face recognition [Ca18] network as our detection D_d and representation systems D_r, respectively. Since the images in the UCCS dataset are very difficult and most of the faces were not detected with the default face detector parameters, we had to lower the detection thresholds to $(0.1, 0.2, 0.2)$. With this setting, most of the faces were detected, but also a large number of background regions were marked as faces.

We detected all faces in all images and extracted the 2048-dimensional deep features of the VGG2 face recognition network for all detected bounding boxes. In total, we obtained 11299 of the 11315 known and 15792 of the 15551 unknown faces,[4] as well as 74962 background detections in the training set. Additionally to the deep features from the training set images of the known and the unknown subjects, we added all background detections of the face detector, which we used as additional unknown samples.[5]

The topology of our Objectosphere adapter network is a simple three-layer fully-connected network with 128 and 64 neurons in the first two layers, and the number of known faces in the UCCS dataset in the last layer. We trained the network with our Objectosphere loss on 90 % of the training data that contained known and unknown subjects, leaving 10 % for validation and ran 1000 training epochs using tensorflow [Ab16]. The training procedure took around 20 minutes on a regular desktop computer with a single NVidia Titan X GPU until convergence on the validation set, which was achieved after around 100 epochs. If more speed is required, the network topology can surely be adapted without a significant loss in accuracy, or more GPU resources could be used.

After training, we extracted the features from our Objectosphere network for all of the known subjects. We enrolled the gallery templates by a simple average of the normalized features so that we had a 64-dimensional template representation G_g of each subject g. Particularly, we also enrolled one gallery template for the unknown faces by computing the average 64-dimensional feature vector over all unknown faces.[6] During testing, for each detected bounding box in the validation set, we used the VGG2 features and the 64-dimensional Objectosphere features as probes.

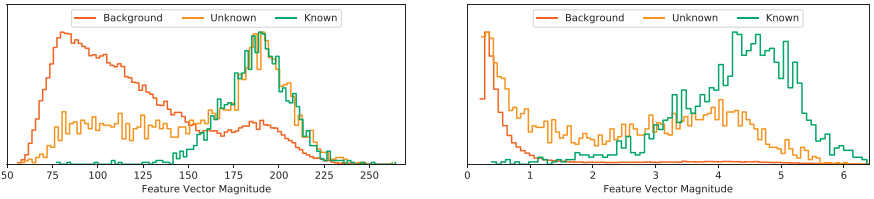
4.3 Deep Feature Magnitudes

One of the goals of Objectosphere is that the deep features $P = D_c(R)$ of unknown samples have a much smaller magnitude than those of known samples. Fig. 3 shows histograms of the feature magnitudes of all probe features of the validation set. When looking at the distribution of the Objectosphere feature magnitudes $||P||$ in Fig. 3(b), we can observe that the background samples are well-separated from the known samples and have very low magnitudes. The known samples are distributed around the desired target magnitude of $\xi = 5$, while the unknown samples have a peak close to 0, but are distributed throughout the range of magnitudes, which might be an effect of badly labeled images in the UCCS

[4] Several faces were detected multiple times and we used all of the detections.

[5] The additional unknown samples provide only a minor improvement. The background detections are to dissimilar to real faces, whereas Objectosphere requires hard negative samples to obtain good results.

[6] This additional template does not change the shape of the DIR in Fig. 4, neither for Objectosphere nor for Softmax. It only reduces the number of false alarms with very low scores and, thus, the plots in the DIR do not extend further to the right. It can be removed in an operational setting that relies on a low FAI.



(a) Original VGG2 features (b) Objectosphere Features

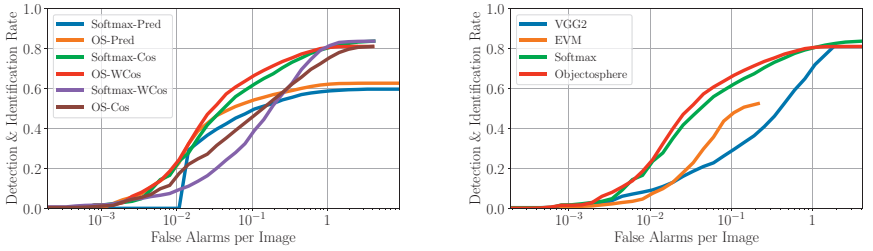
Fig. 3: FEATURE MAGNITUDE HISTOGRAMS. *Histograms of magnitudes of deep features are shown for known subjects, unknown subjects, and background detections, with deep features being extracted (a) from the original VGG2 network and (b) from the Objectosphere network. For visualization purposes, histograms are normalized individually to have the same maximum value.*

dataset, cf. Sec. 4.6. Hence, it is very easy to separate background detections from faces, but a little bit more difficult to separate known from unknown faces.

For comparison, the magnitudes $||R||$ of the original VGG2 features are shown in Fig. 3(a). As expected, the known and unknown samples have similar magnitudes, though many of the unknown samples still have lower values, which we attribute to the fact that many of the unknown faces are of very bad quality. Also, the background detections have relatively low magnitudes, but the overlap with the known samples is large. Hence, the feature magnitude $||R||$ of the original VGG2 network cannot be taken directly as an indicator if features belong to faces or to the background.

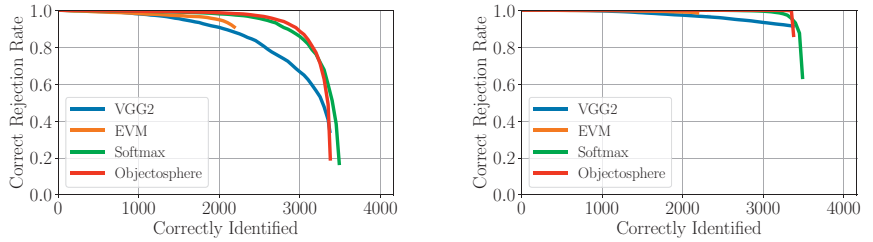
4.4 Softmax vs. Objectosphere

To show the advantage of our Objectosphere training procedure over softmax, we trained a network with identical topology and training strategy with softmax loss, where the negative class contained the same combination of unknown subjects and background detections from the training set. For both networks, we apply three different strategies to identify probe samples. First, we take the network predictions $S_g(D_c(R))$ as similarity values between the probe and all gallery samples. For the softmax-trained network, we ignore the unknown class prediction, but threshold on the predictions of the known classes. For the Objectosphere-trained network, we also obtain the predictions, but we additionally multiply them with the feature magnitude $||P_p||$. As the second alternative, we extract the deep features from both networks, enroll a template G_g for each training subject, and compare template and probe features P_p using the simple cosine similarity. Third, we use the same templates and probes as before, but this time we multiply the cosine distance by the probe feature magnitude (mcos) as in (3). From the DIR plots in Fig. 4(a), we can observe that the extraction of deep features from our networks performs considerably better than the predictions. As anticipated, comparing deep features from the softmax-trained network works far better with the simple cosine similarity rather than the weighted cosine. On the other hand, for Objectosphere the exact opposite is the case, and the Objectosphere network performs considerably better than the softmax-trained network, particularly at lower FAIs.



(a) Softmax vs. Objectosphere (b) State of the Art

Fig. 4: DIR PLOTS. *In (a), DIR curves are generated for two approaches: using the final network prediction as well as enrolling models from deep features and comparing them with two similarity functions; on two different networks: softmax trained with a background class and Objectosphere. In (b), we show the comparison of our two watchlist-adapted networks (Softmax and Objectosphere) with respect to the results of the best participant (EVM) of the face recognition challenge on the UCCS dataset [Gü17b] and the original VGG2 features.*



(a) Unknown Subjects (b) Background Detections

Fig. 5: REJECTION RATES BY TYPE. *Unknown samples are split into (a) unknown subjects and (b) background detections. We plot, how many samples are correctly rejected (i.e., not identified as any known subject) for given thresholds that are based on the number of correctly identified known subjects. Both adapters (Softmax and Objectosphere) significantly improve rejection of unknowns.*

4.5 Comparison to the State of the Art

In order to compare our results to other work that reported on the UCCS dataset, we plot the results of the best participant, who used the extreme value machine (EVM). We generated the DIR curves on the validation set, where we include the VGG2 baseline that was the basis of our networks trained with softmax and Objectosphere, the results of those two networks, and the current state of the art on the UCCS dataset. The resulting DIR curve can be found in Fig. 4(b). Compared to the VGG2 baseline, which is the state of the art [Ca18] on the IARPA Janus Benchmark-A dataset, our Objectosphere improved results drastically, especially at relevant FAI thresholds. Our improved performance might be due to the very different imagery in the UCCS and IJB-A dataset and, in opposition to VGG2, we trained our networks on this type of data. More importantly, we outperform the EVM algorithm, which also trained on the UCCS dataset.

To further investigate the performance of the different systems, we evaluate the number of correctly rejected unknown samples as these samples are disregarded in DIR plots. Basing our score threshold θ on the number of correctly identified known subjects, in Fig. 5 we

(a) Corrected Labels

(b) Incorrect Predictions

Fig. 6: ERROR ANALYSIS. *This figure includes six probe images (left in each triplet) and a selection of two automatically assigned gallery images (right in the triplets, green boxes). In (a) indicated by blue boxes, we present faces for which Objectosphere provided correct labels. As evident, some of these are difficult cases. (b) shows failure cases, indicated by red boxes, where Objectosphere provided an incorrect label.*

plot the percentage of unknown samples that are correctly rejected under this threshold. We separate the unknown samples into unknown subjects and background detections. From Fig. 5(a) we can observe that Objectosphere is able to identify 2000 out of the 4171 known probe faces while rejecting almost all unknown subjects. This is better than all of the related algorithms and particularly much better than the VGG2 baseline. Only the softmax-trained network comes close to our proposed Objectosphere network and exceeds it slightly on the right-hand side of the plot. Checking the rejection performance for background detections in Fig. 5(b) it is clear that Objectosphere is able to reject almost all of the background samples. Even with a threshold that allowed 3000 out of the 4171 known probe faces to be identified correctly, the rejection rate of background detections is very close to 100 %, which is considerably higher than all other algorithms. Even the softmax trained network starts accepting background detections as known subjects. Thus, we can conclude that background detections cannot harm the Objectosphere network anymore, they are successfully rejected by providing very low mcos similarity scores for all subjects in the gallery.

4.6 Failure Analysis

To analyze the errors made by our Objectosphere network, we checked the first 1000 false alarms, i.e., the detected face bounding boxes that were labeled to be unknown in the dataset but that Objectosphere identified as a certain known subject. Out of these error cases, we identified 621 faces, where the automatic label was assigned to a subject that had images from the same time stamp, but for which the assigned subject label and the ground-truth label disagreed. We manually checked those faces by showing a pair of probe face and faces from the anticipated gallery to a human who decided if the pair contains the same face. We selected to use only images with the same time stamp since in this case, additional cues like clothes and neighboring subjects could be included into the manual decision process.

With this process, we found that 573 of the 621 faces actually have a wrong label, most of the faces are labeled as unknown, and our Objectosphere approach was correct. Only in 48 of the 621 cases, our assigned label was actually wrong. Therefore, we assume that the plots shown in Fig. 4 do not reflect the reality, but false alarms actually happen far less

frequently. A few examples are presented in Fig. 6, wherein Fig. 6(a) shows difficult cases where Objectosphere was able to identify the correct subject, which would have been a very hard task for face recognition systems a decade ago. On the other hand, the failure cases displayed in Fig. 6(b) indicate that the network still uses different features than humans would since many of the failure cases are really obvious to humans, at least when the local context around the face is included.

5 Conclusion

Pre-trained networks are widely used for recognition tasks. This is the first paper to demonstrate how they can be adapted to improve open-set recognition in a detection-recognition pipeline. Our approach should adapt any existing state-of-the-art detection and recognition approaches to improve support for rejecting both unknown inputs and background detections.

In this paper, we approached face recognition watchlist as an open-set problem by focusing on decreasing the false alarms of the non-gallery subjects while maintaining/improving the performance of identifying the watchlist subjects. For this purpose we used a novel open-set classification technique called Objectosphere and evaluated its effectiveness on popular face recognition metrics. With the UnConstrained College Students (UCCS) dataset, we employed the largest open-set face dataset to demonstrate this effectiveness. Using the deep features from the VGG2 face recognition network for all the detections, we trained an Objectosphere adapter with background detections and unknown faces and demonstrated the generalization to the test set. We also demonstrate that our adapter network can be used for its representation ability rather than just classification ability it was originally trained for. For probe detections, we showed that enrolling gallery templates and computing similarity scores perform better than using the raw features from a pretrained network, especially when we applied our specifically designed mcos similarity. We found that Objectosphere performs better than the state-of-the-art algorithms that were reported on the UCCS dataset.

The protocol of the UCCS dataset permits training on the gallery, which is known to favor certain types of algorithms [Lu12]. In this paper, we followed that protocol and trained on the gallery, but we are confident that the proposed algorithm would also work in rejecting background detections in protocols that do not allow gallery adaptation. However, since for the representation network all faces from the evaluation dataset are unknown, there is naturally no way of telling apart known from unknown faces without training on the gallery.

Our focus in this paper is on face watchlists which we see as a critically under-studied and socially important problem. While government/system operators may prefer the simplicity of a pre-trained system, this paper shows a significant improvement from using watchlist adaptation. We argue that modern watchlists operators should accept the mild cost of doing watchlist/gallery adaptation to protect the innocent.

References

[Ab16] Abadi, Martín et al.: TensorFlow: A System for Large-scale Machine Learning. In: USENIX Conference on Operating Systems Design and Implementation. 2016. 7

[Ba17] Bansal, Ankan; Nanduri, Anirudh; Castillo, Carlos D.; Ranjan, Rajeev; Chellappa, Rama:

UMDFaces: An Annotated Face Dataset for Training Deep Networks. In: IJCB. 2017. 3

[Ca18] Cao, Qiong; Shen, Li; Xie, Weidi; Parkhi, Omkar M.; Zisserman, Andrew: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: FG. 2018. 3, 7, 9

[Ch16] Chowdhury, Aruni Roy; Lin, Tsung-Yu; Maji, Subhransu; Learned-Miller, Erik: One-to-many Face Recognition with Bilinear CNNs. In: WACV. 2016. 3

[CPC16] Chen, Jun-Cheng; Patel, Vishal M.; Chellappa, Rama: Unconstrained Face Verification using Deep CNN Features. In: WACV. 2016. 3

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. 2019. 3

[DGB18] Dhamija, Akshay Raj; Günther, Manuel; Boult, Terrance E.: Reducing Network Agnostophobia. In: NeurIPS. 2018. 2, 4

[GNH19] Grother, Patrick; Ngan, Mei; Hanaoka, Kayee: Face Recognition Vendor Test (FRVT) Part 2: Identification. Technical report, NIST, 2019. 1

[Gü17a] Günther, Manuel; Cruz, Steve; Rudd, Ethan M.; Boult, Terrance E.: Toward Open-Set Face Recognition. In: CVPR Workshops. 2017. 5

[Gü17b] Günther, Manuel et al.: Unconstrained Face Detection and Open-Set Face Recognition Challenge. In: IJCB. 2017. 3, 6, 9

[Kl15] Klare, Brendan et al.: Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A. In: CVPR. 2015. 3

[Lu12] Lui, Yui Man et al.: Preliminary Studies on the Good, the Bad, and the Ugly Face Recognition Challenge Problem. In: CVPR Workshops. 2012. 2, 11

[Ma18] Maze, Brianna et al.: IARPA Janus Benchmark - C: Face Dataset and Protocol. In: ICB. 2018. 3

[O'18] O'Toole, Alice J. et al.: Face Space Representations in Deep Convolutional Neural Networks. Trends in Cognitive Sciences, 2018. 3

[PGM11] Phillips, P. Jonathon; Grother, Patrick; Micheals, Ross: Evaluation Methods in Face Recognition. In: Handbook of Face Recognition. Springer, 2nd edition, 2011. 5, 6

[PVZ15] Parkhi, Omkar M.; Vedaldi, Andrea; Zisserman, Andrew: Deep Face Recognition. In: BMVC. 2015. 3

[Ro17] Romm, Tony: Amazon's facial-recognition tool misidentified 28 lawmakers as people arrested for a crime, study finds. Washington Post, July 2017. Retrieved from http://www.washingtonpost.com. 1

[Sa16] Sankaranarayanan, Swami; Alavi, Azadeh; Castillo, Carlos D.; Chellappa, Rama: Triplet Probabilistic Embedding for Face Verification and Clustering. In: BTAS. 2016. 3

[SB13] Sapkota, Archana; Boult, Terrance E.: Large Scale Unconstrained Open Set Face Database. In: BTAS. 2013. 3

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: CVPR. 2015. 3

[Wh17] Whitelam, Cameron et al.: IARPA Janus Benchmark-B Face Dataset. In: CVPR Workshops. 2017. 3

[Zh16] Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y.: Joint Face Detection and Alignment using Multitask Cascaded Convolutional Networks. Signal Processing Letters, 2016. 7

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 33

Fisher Vector Encoding of Dense-BSIF Features for Unknown Face Presentation Attack Detection

Lázaro J. González-Soler[1], Marta Gomez-Barrero[2], Christoph Busch[1]

Abstract: The task of determining whether a sample stems from a real subject (i.e, it is a bona fide presentation) or it comes from an artificial replica (i.e., it is an attack presentation) is a mandatory requirement for biometric capture devices, which has received a lot of attention in the recent past. Nowadays, most face Presentation Attack Detection (PAD) approaches have reported a good detection performance when they are evaluated on known Presentation Attack Instruments (PAIs) and acquisition conditions, in contrast to more challenging scenarios where unknown attacks are included in the evaluation. For those more realistic scenarios, the existing approaches are in many cases unable to detect unknown PAI species. In this work, we introduce a new feature space based on Fisher vectors, computed from compact Binarised Statistical Image Features (BSIF) histograms, which allows finding semantic feature subsets from known samples in order to enhance the detection of unknown attacks. This new representation, evaluated over three freely available facial databases, shows promising results in the top state-of-the-art: a BPCER100 under 17% together with a AUC over 98% can be achieved in the presence of unknown attacks.

Keywords: Presentation attack detection, probabilistic visual vocabulary, common feature space, unknown attacks, face.

1 Introduction

The deployment of biometric systems has increased over the last decades. In spite of their advantages, facial recognition systems are also vulnerable to Presentation Attacks (PAs): with the broad development experienced by social networks, an attacker can easily download a photo or video of a given person, thereby gaining access to several applications in which face recognition systems are commonly deployed. Moreover, the recent advances in creating synthetic videos, or deep fakes, also poses a serious threat [To20].

In order to address those concerns, several Software-based face Presentation Attack Detection (PAD) methods have been proposed. In general, many PAD approaches have reported a high detection performance for identifying Presentation Attack Instruments (PAIs) when both the attack type and acquisition conditions are known a priori (i.e., known attacks scenario). In contrast, a rather limited number of works have addressed so far a more realistic and challenging scenario where the PAI species in the test set remain unknown in the training set (i.e., unknown attacks). Back in 2013, de Freitas Pereira *et al.* [Fr13] already reported poor generalisation capabilities to unknown attacks of state-of-the-art face PAD methods based on local binary patterns (LBP) and support vector machines (SVMs). In

[1] dasec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany,
{lazaro-janier.gonzalez-soler;christoph.busch}@h-da.de
[2] Hochschule Ansbach, Germany, marta.gomez-barrero@hs-ansbach.de

particular, the error rates increased by at least 100%. Motivated by these findings, Arashloo *et al.* [AKC17] experimented over several unknown attack scenarios and concluded that anomaly detection approaches trained only on bona fide data can reach a detection performance comparable to two-class classifiers. However, the results are reported only in terms of the area under the Receiving Operating Characteristic curve (AUC), thus lacking a proper quantitative analysis in line with the ISO/IEC 30107-3 standard on biometric PAD [IS17].

More recently, Nikisins *et al.* [Ni18] showed how a one-class Gaussian Mixture Model (GMM) can outperform two-class classifiers, depending on the PAI species included in the test set. The experimental evaluation reported an error rate increase with respect to the known scenario and two-class classifiers. Following the same anomaly detection paradigm, Xiong and AbdAlmageed studied in [XA18] the detection performance of one-class SVMs and autoencoders in combination with LBP descriptors for PAD purposes. In most of the scenarios tested, the detection rates increased with respect to common two-class classifiers. Finally, Liu *et al.* also analysed in [Li19] the performance of a Deep Tree Network (DTN) for facial PAD, which clusters the PAI species into semantic sub-groups. Over a new database comprising 13 PAI species, and following a leave-one-out testing protocol, an average D-EER of 16% is achieved, which is still above the state-of-the-art for known attacks.

To tackle those open issues with unknown attacks, we focus on a different approach which has already shown remarkable results in cross-sensor and unknown attacks scenarios for fingerprint PAD [Go19a]. Gonzalez-Soler *et al.* proposed in [Go19a] a combination of local feature descriptors and global feature encoding to model a new feature space in which the generalisation capabilities of the PAD module are enhanced. In fact, this approach achieved the best detection accuracy in the LivDet 2019 competition [Or19]. Whereas some keypoint based descriptors such as SIFT and SURF have shown to be an appropriate choice for fingerprint samples [Go19a, Go19b], in which minutiae can be regarded as landmarks within the image, we anticipate that for facial images the textural information is more relevant than the geometric information related to facial landmarks. Therefore, we propose a new face PAD approach, which encodes accurate and compact dense Binarized Statistical Image Features (dense-BSIF), extracted from local patches of the facial image, and projects them into a new feature space with Fisher vectors. By assuming that the unknown attacks share more texture, shape and appearance features with known PAIs than with those BP samples, this FV representation allows in turn a definition of semantic subgroups from known samples to tackle the aforementioned issues on PAD generalisation to unknown attacks. In order to validate the detection capabilities of the proposal, a thorough evaluation compliant with the ISO/IEC 30107-3 standard on biometric PAD [IS17] is also carried out over three well-established databases: CASIA Face Anti-Spoofing [Zh12], REPLAY-ATTACK [CAM12] and REPLAY-MOBILE [Co16].

The remainder of this paper is organised as follows. The proposed PAD method is presented in Sect. 2. Sect. 3 describes the experimental protocol and presents the results. Finally, conclusions and future work directions are presented in Sect. 4.

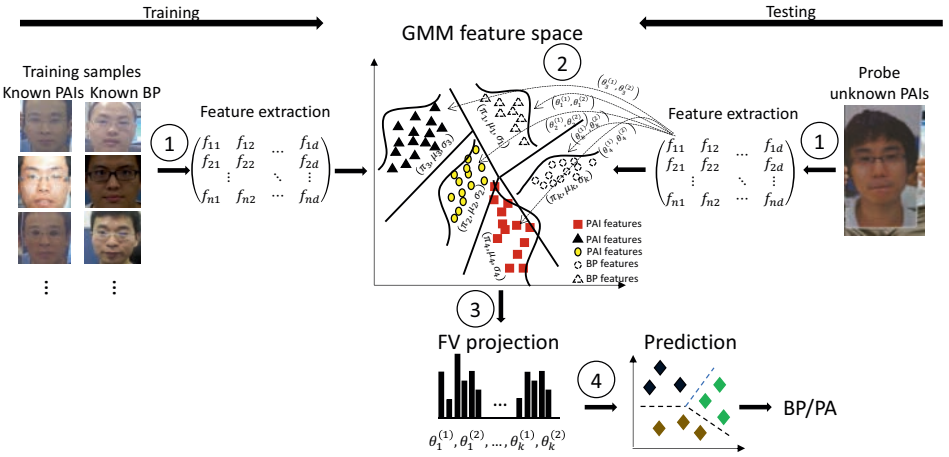


Fig. 1: Face PAD approach overview which comprises three steps: *i*) BSIF descriptors are densely extracted; *ii*) the BSIF data distribution is subsequently learnt by training an unsupervised Gaussian Mixture Model (GMM) from known samples; *iii*) an unknown sample at hand is then encoded by computing the gradient among their BSIF components and parameters obtained by the GMM; and *iv*) the final features are finally classified using a linear SVM.

2 Proposed Method

We build upon the fingerprint PAD approach presented in [Go19a]. Fig. 1 shows an overview of the proposed PAD approach, which consists of four main steps: (1) dense-BSIF histograms are extracted from a face sample, which has been detected by the Viola and Jones method [VJ04]; (2) our new feature space is built by learning an unsupervised Gaussian Mixture Model (GMM) model from the aforementioned features; (3) the final descriptors are subsequently encoded by computing the differences of first- and second-order statistics with respect to the learned model parameters; and (4) a bona fide presentation (BP) or presentation attack (PA) decision is taken by a linear SVM.

2.1 Dense-BSIF Descriptors

Usually, PAIs include artefacts (e.g. acute edges in the cut eyes of the CASIA images [Zh12]) which can be successfully detected by the quantization of filtered features. BSIF [KR12] is a local image descriptor computed by binarising the responses of a given image to a set of pre-learned filters to obtain a statistically meaningful representation of the data. More specifically, let X be an image patch of size $l \times l$ and $W = \{W_1, \ldots, W_N\}$ a set of linear filters of the same size as X. Then, we compute binarised responses b_n:

$$b_n = \begin{cases} 1 & \sum_{u,v} W_n(u,v)X(u,v) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

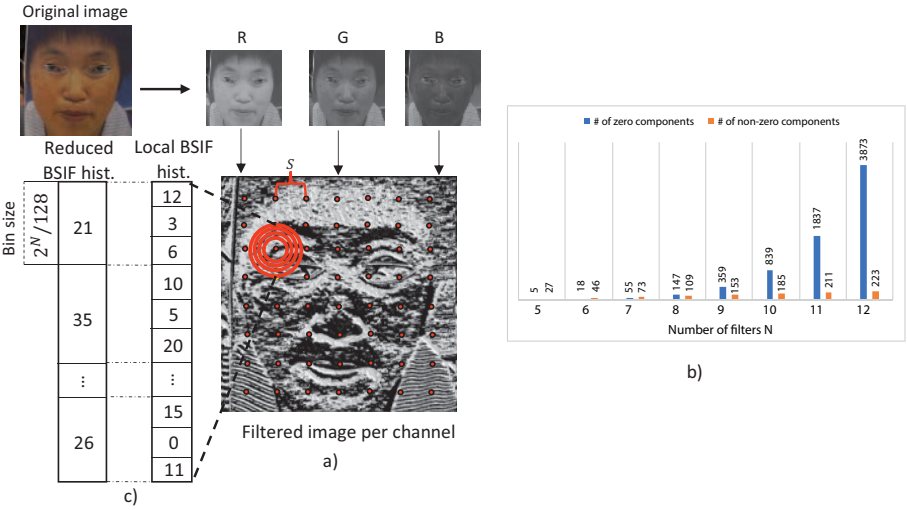


Fig. 2: BSIF feature extraction. a) BSIF histograms are densely computed at fixed points on a regular grid with a stride of S pixels, b) average number of zero and non-zero components of dense-BSIF histograms for different numbers of filters N, and c) a reduction example where a local BSIF histogram of size $2^N = 512$ is represented as a 128-component vector.

All the filter responses b_n are subsequently stacked to form a bit string $\mathbf{b}$ with size N for each pixel. Subsequently, $\mathbf{b}$ is transformed to a decimal value, and then a 2^N histogram for X is computed. In our work, 60 filter sets with different sizes $l = \{3,5,7,9,11,13,15,17\}$ and number of filters $N = \{5,6,7,8,9,10,11,12\}$ were obtained from [KR12].

Now, given that artefacts can be detected at any point in the image, not only at relevant facial landmarks, BSIF-histograms are densely extracted over a regular grid with a fixed stride S of 3. Furthermore, in order to capture local and global information of the artefacts produced in the fabrication of the PAIs, histograms are computed over four circular patches with different radii $r = \{4,6,8,10\}$, as depicted in Fig. 2a). Therefore, each point in the grid is represented by four dense-BSIF histograms.

In addition, we have observed that the histograms become sparse vectors as the number of linear filters N increases. Therefore, we computed the number of zero and non-zero components per number of filters over the CASIA Face Anti-Spoofing database [Zh12] in Fig. 2b), and noted that the number of non-zero components remains under 223 in all cases, having an average value of 128. We will thus represent each 2^N BSIF histogram as a 128-component vector by summing the elements for each sequential $2^N/128$ sub-set in the original histogram (see an example in Fig 2c)). This representation reduces the storage requirements down to 12.5% for N = 10 or 3.1% for N = 12.

2.2 Fisher Vector Encoding

The Fisher vector (FV) feature encoding approach transforms local features into an new feature space based on the parameters learnt by a generative model, known as visual vocabulary [PSM10]. This representation describes how the distribution of these local descriptors extracted from unknown PAIs differs from the known PAI distribution previously learned. In particular, a Gaussian Mixture Model (GMM) with K-components, which is represented by their mixture weights (π_k), means (μ_k), and covariance matrices (σ_k), with $k = 1, \ldots, K$, allows discovering semantic sub-groups from known PAIs and BP samples, which could successfully enhance the detection of unknown attacks. In order to compute those semantic groups, the compact dense-BSIF descriptors (see Sect. 2.1) are decorrelated using Principal Component Analysis (PCA) [Je12], hence reducing their size to $d = 64$ components while retaining 98% of the variance. Then, the average first-order and second-order differences between the given decorrelated features and each semantic sub-group are computed, thereby obtaining a $2Kd$ dimensional vector. For the GMM computation, we selected $K = 1024$, since it allows representing a more complex structure from data while preserving low computational requirements. Therefore, each facial image is finally represented by a vector with size $2Kd = 2 \cdot 1024 \cdot 64 = 131072$.

3 Experimental Evaluation

3.1 Experimental Protocol

The experimental evaluation was conducted over three well-established databases for facial PAD, whose images were captured with different resolutions and several acquisition conditions: CASIA Face Anti-Spoofing [Zh12], REPLAY-ATTACK [CAM12], and REPLAY-MOBILE [Co16].

The experimental protocol aims to address the following goals: *i*) analyse the impact of different BSIF filter configurations in terms of number of filters and filter's size on the detection performance of our PAD approach, *ii*) study its detection performance for each RGB colour component under known and unknown attacks, and *iii*) benchmark the detection performance of our proposed PAD approach against the top state-of-the-art. Keeping these goals in mind, we defined two different scenarios:

- *Known-attacks*: which includes an analysis of all PAI species. In all cases, PAI species for testing are included in the training set, as described in [Zh12].
- *Unknown-attacks*, in which the PAI species used for testing are not incorporated in the training set. In particular, we consider using the leave-one-out testing protocol explained in [AKC17], in which a PAI is only evaluated whilst the remaining PAI species are employed for training.

Finally, the experimental evaluation is conducted in compliance with the international metrics of ISO/IEC 30107-3 [IS17]: *i*) Attack Presentation Classification Error Rate (APCER)

Tab. 1: Benchmark in terms of D-EER(%) of our PAD approach per colour component and the whole RGB colour space against the top state-of-the-art methods.

Method	Warped	Cut	Video	Overall
BSIF + SVM [RB14]	-	-	-	10.21
MBSIF-TOP [AKC15]	1.40	10.10	4.30	7.20
Texture fusion [BKH18]	-	-	-	4.60
ResNet-15 on 3D [Gu19]	-	-	-	2.22
shallowCNN-LE [QDN19]	-	-	-	4.00
SPMT + SSD [So19]	0.35	0.20	0.03	0.04
Hybrid residual DL [MH19]	-	-	-	0.02
Proposed Method (R)	1.42 ± 1.04	2.20 ± 0.83	0.28 ± 0.56	2.92 ± 0.93
Proposed Method (G)	1.44 ± 1.11	2.22 ± 0.79	0.50 ± 0.72	2.52 ± 1.18
Proposed Method (B)	1.59 ± 1.01	2.78 ± 1.27	0.59 ± 0.77	2.53 ± 0.99
Proposed Method (RGB)	1.20 ± 0.77	1.74 ± 0.75	0.30 ± 0.54	1.79 ± 0.82

which is defined as the proportion of attack presentations wrongly classified as bona fide presentations, and *ii*) Bona Fide Presentation Classification Error Rate (BPCER) which is the proportion of bona fide presentations missclassified as attack presentations. We therefore report: *i*) the Detection Error Trade-off (DET) curves between both APCER and BPCER, *ii*) the BPCER values for several security thresholds (BPCER10, BPCER20 and BPCER100), and *iii*) the Detection Equal Error Rate (D-EER), which are defined as the error rate value at the operating point where APCER = BPCER.

3.2 Results and Discussion

3.2.1 Known Attacks

First, we need to find the optimal configuration of our proposed method in terms of the filter size l, the number of BSIF filters N, and the best performing RGB component. To that end, we compute error rates for each of sixty filter configuration and report in Table. 1 the mean and standard deviation (std) of the D-EER achieved by each particular RGB component and their fusion (i.e., RGB) over the Attack test and Overall test protocol in the CASIA Face Anti-Spoofing database [Zh12]. As it could be expected, the entire RGB space reports on average the best detection performance, since it fuses the information of the three channels. In addition, it reports for $N = 11$ filters of size $l = 11$ a minimum D-EER of 0.0% for warped, 1.11% for cut photo attacks, 0.0% for video replay attacks, and 0.37% for the overall test, which achieve the top state-of-the-art.

On the other hand, it may also be observed that among the three individual RGB colour components, the R channel appears to be the one which contains the most discriminative features for the PAD task. Specifically, it achieves for $N = 9$ filters of size $l = 6$, a D-EER of 0.0% for warped photo attacks, 2.22% for cut photo attacks, 0.0% for video replay attacks,

Tab. 2: D-EER(%) values under the *Unknown-attacks* protocol for RGB and benchmark, in terms of AUC(%), for the best unknown attack setting (i.e., $N = 10$ filters of size $l = 9$), against the top state-of-the-art approaches.

	CASIA			**REPLAY-ATTACK**			**REPLAY-MOBILE**		
	Cut	Warped	Video	Digital	Printed	Video	Digital	Printed	Video
OC-SVM$_{RGB}$+BSIF [AKC17]	60.7	95.9	70.7	88.1	73.7	84.3	-	-	-
NN+LBP [XA18]	88.4	79.9	94.2	95.2	78.9	99.8	-	-	-
DTL [Li19]	97.3	97.5	90.0	99.9	99.6	99.9	-	-	-
our proposal (AUC)	**99.6**	**97.9**	**99.9**	**100**	**99.9**	**100**	**100**	**100**	**100**
our proposal (D-EER)	4.11 ± 1.99	6.15 ± 2.42	1.37 ± 1.60	0.00 ± 0.00	1.35 ± 1.73	0.00 ± 0.00	0.00 ± 0.00	0.34 ± 0.63	0.02 ± 0.12

and 0.37% for the overall test, which are almost the same minimum D-EER reported by the entire RGB colour space.

In order to validate the detection performance of our PAD approach under different RGB configurations, we select the non-parametric Mann-Whitney test with a 95% of confidence and verify the statistical significance of error rates reported by each RGB component. To that end, we define as null and alternative hypothesis:

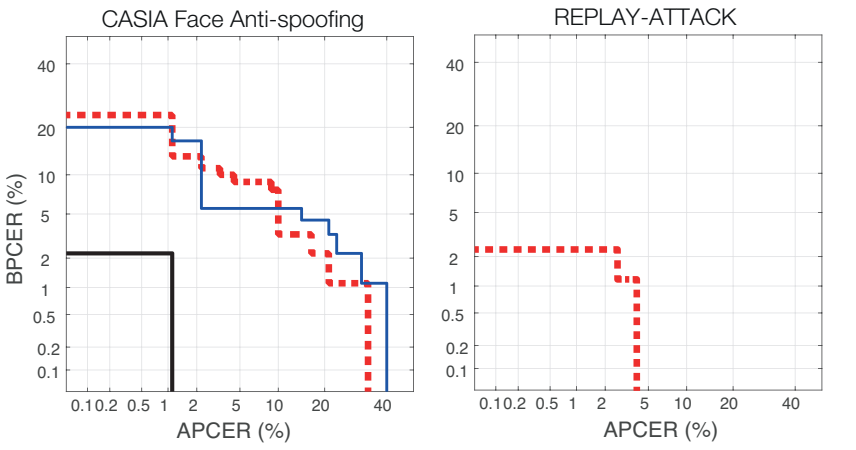
- H_0: two colour components provide the same discriminative information for PAD.
- H_1: two colour components do not provide the same discriminative information for PAD.

Then, an all-against-all comparison per database is performed. As a result of this test, we do confirm that error rates for the three RGB colour components stem from the same population, thereby showing the same discriminative information for facial PAD. In contrast, results for the entire RGB colour space claim to be statistically better than each particular component. Therefore, we do confirm that even if each individual RGB component is correlated with each other, the entire RGB colour space includes some additional information which allows learning more discriminative features for facial PAD.

3.2.2 Unknown Attacks

As it was mentioned in Sect. 1, one of the goals of our work is to successfully identify unknown PAIs. To that end, a set of experiments is carried out over the three selected databases following the leave-one-out protocol described in [AKC17]: two PAI species are included in the training set, and the last one in the test set. Table 2 reports the corresponding D-EER values, a benchmark against the top state-of-the-art PAD approaches and the complete DET curves are depicted in Fig. 3.

Taking a look at Table 2, we can observe how error rates for each particular unknown PAIs with respect to the corresponding known attack are multiplied by a factor of 2.36% for cut photo attacks, 4.57% for video replay attacks and 5.13% for warped photo attacks, respectively, for the CASIA database. However, for a fixed filter configuration, it should be noted



Train vs test	BPCER10	BPCER20	BPCER100
warped + video vs cut	5.56	5.56	16.7
video + cut vs warped	3.33	8.89	13.3
cut + warped vs video	0.00	0.00	0,00

Train vs test	BPCER10	BPCER20	BPCER100
printed + video vs digital	0.00	0.00	0,00
digital + video vs printed	0.00	0.00	2.50
printed + digital vs video	0.00	0.00	0,00

Fig. 3: *Unknown-attacks* DET curves and BPCER(%) values over the leave-one-out protocol for the CASIA and REPLAY-ATTACK databases. The REPLAY-MOBILE database reports a BPCER = 0.0% for any APCER.

that our proposed approach outperforms, in terms of AUC, the top state-of-the-art for all attack types. Since we lack a proper quantitative analysis of the top state-of-the-art methods in compliance with the ISO/IEC 30107-3 standard on biometric PAD [IS17], we can only establish a benchmark in terms of AUC. For the REPLAY's databases, a higher detection performance outperforming the top-state-of-the-art techniques can be observed: an AUC of almost 100% for the entire set of attacks confirms the soundness of our proposed method to identify PAIs stemming from this challenging scenario.

In addition, Fig. 3 confirms the detection performance showed by our approach: a low BPCER100 of 0.0%, 0.0%, 13.3% and 16.7% are achieved by video replay, digital, cut photo and warped photo attacks, respectively for a high security threshold (i.e., APCER of 1%). This in turn yields, in this challenging scenario, a secure (only one in 100 attacks are not detected) and convenient (zero to seventeen in 100 bona fide presentation attempts are rejected) system.

Finally, a t-SNE visualisation in Fig. 4 of BP and PA samples are in the CASIA database confirms the aforementioned hypotheses, which state that the PAIs share more texture, shape and appearance features with known PAIs than with those BP samples. Whereas the FV representations of attack presentations (blue, red and yellow) are separated of the bona fide presentations (green spots), they are close to each other. However, we can also observe that some PAIs, such as warped (yellow) and cut photo attacks (blue), still overlap with

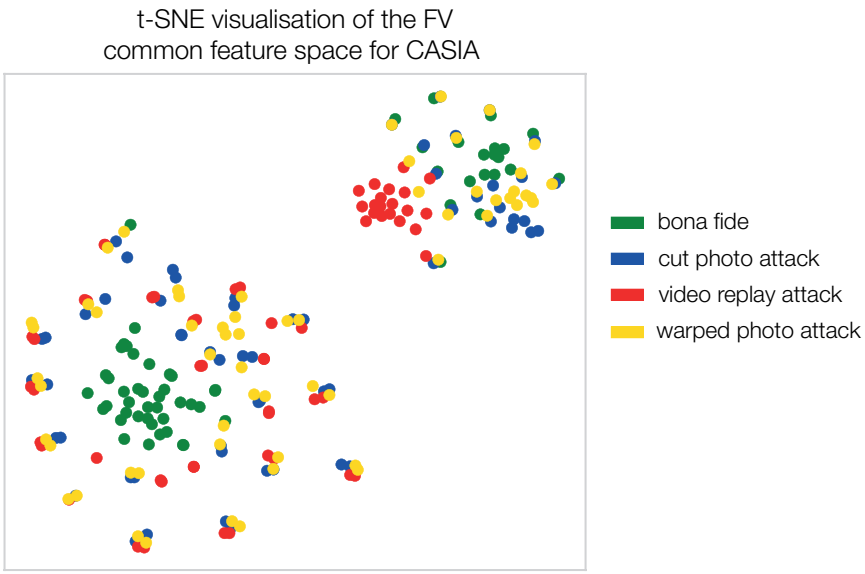


Fig. 4: t-SNE visualisation for BP vs. PA samples in the CASIA Face Anti-spoofing database.

BP samples, thereby indicating that the data distribution learned by a GMM model using the BSIF features needs to be improved in order to get a better detection performance.

4 Conclusions

In this work, a new face PAD approach to generalise to unknown attacks was proposed. In essence, it projects compact dense-BSIF descriptors into a new feature space, which allows discovering semantic feature sub-groups from known samples in order to improve the PAD generalisation capability. In addition, a new strategy for computing compact dense-BSIF histograms was presented, which can be applied to any other texture recognition application. In more details, a reduction down to 95% in the feature vector length can be achieved with no significant impact on the recognition accuracy but strongly reducing the time required for PAD analysis. The experimental evaluation over three freely available databases confirmed the soundness of our proposal for detecting both known and unknown PAIs. Specifically, experimental results indicated the statistical advantage of the entire RGB colour space with respect to each of its particular components, thereby resulting in a minimum D-EER of 0.37% for known attack detection. Finally, BPCER100 in a range of 0.0% to 17% for unknown attack detection, which outperform the top state-of-the-art, showed that our PAD approach is able to yield a secure and convenient system under that challenging scenario. As future work, we plan to evaluate our proposal on larger databases for other colour spaces, which have shown to be superior in terms of detection performance. In addition, a more thorough analysis on larger databases including a higher number of different PAI species will be carried out.

Acknowledgements

This research work has been funded by the DFG-ANR RESPECT Project (406880674) 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

[AKC15] Arashloo, S. R.; Kittler, J.; Christmas, W.: Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features. IEEE Trans. on Information Forensics and Security, 10(11):2396–2407, 2015.

[AKC17] Arashloo, S. R.; Kittler, J.; Christmas, W.: An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. IEEE Access, 5:13868–13882, 2017.

[BKH18] Boulkenafet, Z.; Komulainen, J.; Hadid, A.: On the generalization of color texture-based face anti-spoofing. Image and Vision Computing, 77:1–9, 2018.

[CAM12] Chingovska, I.; Anjos, A.; Marcel, S.: On the Effectiveness of Local Binary Patterns in Face Anti-spoofing. 2012.

[Co16] Costa-Pazo, A.; Bhattacharjee, S.; Vazquez-Fernandez, E.; Marcel, S.: The REPLAY-MOBILE Face Presentation-Attack Database. In: Proc. Intl. Conf. on Biometrics Special Interests Group (BIOSIG). 2016.

[Fr13] de Freitas Pereira, T.; Anjos, A.; Martino, J. De; Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: Proc. Int. Conf. on Biometrics (ICB). pp. 1–8, 2013.

[Go19a] González-Soler, L. J.; Gomez-Barrero, M.; Chang, L.; Pérez-Suárez, A.; Busch, C.: Fingerprint Presentation Attack Detection Based on Local Features Encoding for Unknown Attacks. arXiv preprint arXiv:1908.10163, 2019.

[Go19b] González-Soler, L. J.; Gomez-Barrero, M.; Chang, L.; Pérez-Suárez, A.; Busch, C.: On the Impact of Different Fabrication Materials on Fingerprint Presentation Attack Detection. In: Proc. Intl. Conf. on Biometrics (ICB). 2019.

[Gu19] Guo, J.; Zhu, X.; Xiao, J.; Lei, Z.; Wan, G.; Li, S. Z.: Improving Face Anti-Spoofing by 3D Virtual Synthesis. In: Proc. Intl. Conf. on Biometrics (ICB). 2019.

[IS17] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC FDIS 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. International Organization for Standardization, 2017.

[Je12] Jegou, H.; Perronnin, F.; Douze, M.; Sánchez, J.; Perez, P.; Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Trans. on Pattern Analysis and Machine Intelligence, 34(9):1704–1716, 2012.

[KR12] Kannala, J.; Rahtu, E.: BSIF: Binarized statistical image features. In: 2012 21st Intl. Conf. on Pattern Recognition (ICPR). pp. 1363–1366, 2012.

[Li19] Liu, Yaojie; Stehouwer, Joel; Jourabloo, Amin; Liu, Xiaoming: Deep Tree Learning for Zero-shot Face Anti-Spoofing. In: Proc. Conf. on Computer Vision and Pattern Recognition. pp. 4680–4689, 2019.

[MH19] Muhammad, U.; Hadid, A.: Face Anti-spoofing using Hybrid Residual Learning Framework. In: Proc. Intl. Conf. on Biometrics (ICB). 2019.

[Ni18] Nikisins, O.; Mohammadi, A.; Anjos, A.; Marcel, S.: On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In: Proc. Int. Conf. on Biometrics (ICB). pp. 75–81, 2018.

[Or19] Orrù, G.; Casula, R.; Tuveri, P.; Bazzoni, C.; Dessalvi, G.; Micheletto, M.; Ghiani, L.; Marcialis, G. L.: Livdet in action-fingerprint liveness detection competition 2019. In: Proc. Intl. Conf. on Biometrics (ICB). IEEE, pp. 1–6, 2019.

[PSM10] Perronnin, F.; Sánchez, J.; Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. European Conf. on Computer Vision (ECCV). pp. 143–156, 2010.

[QDN19] Qu, X.; Dong, J.; Niu, S.: shallowCNN-LE: A shallow CNN with Laplacian Embedding for face anti-spoofing. In: Intl. Conf. on Automatic Face & Gesture Recognition. pp. 1–8, 2019.

[RB14] Raghavendra, R.; Busch, C.: Presentation attack detection algorithm for face and iris biometrics. In: Proc. European Signal Processing Conf. (EUSIPCO). pp. 1387–1391, 2014.

[So19] Song, X.; Zhao, X.; Fang, L.; Lin, T.: Discriminative representation combinations for accurate face spoofing detection. Pattern Recognition, 85:220–231, 2019.

[To20] Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J.: DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. arXiv preprint arXiv:2001.00179, 2020.

[VJ04] Viola, P.; Jones, M. J.: Robust real-time face detection. Intl. Journal of Computer Vision, 57(2):137–154, 2004.

[XA18] Xiong, F.; AbdAlmageed, W.: Unknown presentation attack detection with face RGB images. In: Proc. Int. Conf. on Biometrics Theory, Applications and Systems (BTAS). pp. 1–9, 2018.

[Zh12] Zhang, Z.; Yan, J.; Liu, S.; Lei, Z.; Yi, D.; Li, S. Z: A face antispoofing database with diverse attacks. In: Proc. Intl. Conf. on Biometrics (ICB). pp. 26–31, 2012.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 45

Touchless Fingerprint Sample Quality: Prerequisites for the Applicability of NFIQ2.0

Jannis Priesnitz[1,2], Christian Rathgeb[1], Nicolas Buchmann[2], Christoph Busch[1]

Abstract: The impact of fingerprint sample quality on biometric performance is undisputed. For touch-based fingerprint data, the effectiveness of the NFIQ2.0 quality estimation method is well documented in scientific literature. Due to the increasing use of touchless fingerprint recognition systems a thorough investigation of the usefulness of the NFIQ2.0 for touchless fingerprint data is of interest.

In this work, we investigate whether NFIQ2.0 quality scores are predictive of error rates associated with the biometric performance of touchless fingerprint recognition. For this purpose, we propose a touchless fingerprint preprocessing that favours NFIQ2.0 quality estimation which has been designed for touch-based fingerprint data. Comparisons are made between NFIQ2.0 score distributions obtained from touch-based and touchless fingerprint data of the publicly available FVC06, MCYT, PolyU, and ISPFDv1 databases. Further, the predictive power regarding biometric performance is evaluated in terms of Error-versus-Reject Curves (ERCs) using an open source fingerprint recognition system. Under constrained capture conditions NFIQ2.0 is found to be an effective tool for touchless fingerprint quality estimation if an adequate preprocessing is applied.

Keywords: Biometrics, Fingerprint, Touchless Fingerprint, Sample Quality.

1 Introduction

In the past decade, many research efforts have been devoted to robust fingerprint quality estimation, for comprehensive surveys the reader is referred to [OŠB16, BVS14]. It is generally conceded that fingerprint quality assessment is vital to achieve competitive recognition accuracy, *i.e.* quality estimation serves as a predictor of biometric performance. NIST published the first open algorithm for finger image quality assessment which is referred to as NIST Fingerprint Image Quality (NFIQ) in 2004 [TWW04]. Its improved successor, NFIQ2.0 [NI], represents a well-established tool for quality estimation which is used in many operational fingerprint recognition systems. NFIQ2.0 has been specifically designed to assess the quality of fingerprints acquired by touch-based sensors which are optical capture devices and provide fingerprint images of 500dpi spacial resolution.

Touchless fingerprint recognition represents a rapidly growing field of research, for overviews of published scientific literature the reader is referred to [La14, Ma17]. A comparison of a touch-based and touchless fingerprint representation is depicted in Figure 1. In touchless fingerprint recognition methods, effective quality control is of utmost importance as

[1] da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Schöfferstraße 9, 64295 Darmstadt, Germany, firstname.lastname@h-da.de
[2] Freie Universität Berlin, Takustraße 9, 14195 Berlin, Germany, firstname.lastname@fu-berlin.de

(a) touch-based (b) touchless

Fig. 1: Touch-based and touchless fingerprint representations of a single finger: touch-based fingerprint acquired with a Crossmatch Guardian 200 (left); touchless fingerprint image captured with a Samsung Galaxy S8 smartphone (right).

numerous factors may negatively impact fingerprint quality.In many proposed touchless systems captured fingerprint images are pre-processed in a way that these resemble properties of touch-based fingerprint imagery, *e.g.* in terms of contrast or image resolution. This entails two major advantages: on the one hand, sub-systems of touch-based recognition systems for quality control, feature extraction, and comparison can be maintained; on the other hand, acquired touchless imagery can be compared to legacy data.

Focusing on touchless fingerprint recognition, some dedicated quality estimation methods have been proposed, *e.g.* [YLB13, LPS10, Li13]. Labati *et al.* [LPS10] showed that a direct application of NFIQ (version 1) to touchless fingerprint images generally yields low quality scores. The authors conclude that NFIQ1.0 is not usable for touchless fingerprint imagery. In contrast, Salum *et al.* [Sa17] showcased that good NFIQ1.0 scores can be obtained in case touchless fingerprints are pre-processed adequately. To the best of the authors' knowledge the applicability of NFIQ2.0 to touchless fingerprint data has not been investigated.

This work investigates the usefulness of NFIQ2.0 in the context of touchless fingerprint recognition. First, the NFIQ2.0 score distributions of well-known touch-based fingerprint databases and publicly available touchless fingerprint databases are compared. For this purpose, a pre-processing pipeline is proposed which favours the extraction of NFIQ2.0 scores from touchless fingerprints. Further, the predictive power of NFIQ2.0 on touchless fingerprint data is estimated in terms of Error-versus-Reject Curves (ERCs) as suggested by Grother and Tabassi [GT07]. Based on biometric performance rates, quality score distributions, and shapes of ERCs different conclusions w.r.t. the applicability of NFIQ2.0 for touchless fingerprint data are reached.

Tab. 1: Overiew of used fingerprint databases. The DPI value is listed if it is specified in the database description.

Database	Subset	Type	Sensor	Color	Resolution	Instances	Samples
FVC06	DB2-A	touch-based	optical	grayscale	400×560 (569 dpi)	140	1,680
	DB3-A	touch-based	thermal sweeping	grayscale	400×500 (500 dpi)	140	1,680
	DB4-A	synthetic	–	grayscale	288×384	140	1,680
MCYT	dp (Digital Persona)	touch-based	optical	grayscale	256×400 (500 dpi)	3,300	39,600
	pb (Precise Biometrics)	touch-based	capacitive	grayscale	300×300 (500 dpi)	3,300	39,600
PolyU	CB-S1 (contact-based session 1)	touch-based	optical	grayscale	328×356	336	2,016
	CB-S2 (contact-based session 2)	touch-based	optical	grayscale	328×356	160	960
	CL-S1 (contactless session 1)	touchless	digital camera, LED light	RGB	1,400×900	336	2,016
	CL-S2 (contactless session 2)	touchless	digital camera, LED light	RGB	1,400×900	160	960
ISPFDv1	LS (live scan)	touch-based	optical	grayscale	544×253 (250 dpi)	128	1,024
	NI (natural indoor)	touchless	Apple iPhone 5	RGB	3,264×2,448	128	1,024
	NO (natural outdoor)	touchless	Apple iPhone 5	RGB	3,264×2,448	128	1,024
	WI (white indoor)	touchless	Apple iPhone 5	RGB	3,264×2,448	128	1,024
	W0 (white outdoor)	touchless	Apple iPhone 5	RGB	3,264×2,448	128	1,024

This paper is organized as follows: Section 2 summarizes the used fingerprint databases. In Section 3 the proposed evaluation pipeline is described in detail. Experimental results are presented in Section 4. Finally, Section 5 concludes.

2 Databases

We employ four different databases, which comprise touch-based as well as subsets of touchless fingerprint images. The use of touch-based fingerprint databases allows for a detailed comparison of NFIQ2.0 quality scores as well as their predictive power w.r.t. biometric performance on touchless and touch-based data. Used databases and their properties are listed in Table 1 and briefly summarized as follows:

- *FVC06* [Ca07]: the database of the fourth international Fingerprint Verification Competition (FVC), containing four disjoint fingerprint subsets. The first three subsets are each collected with a different touch-based sensor while the fourth database is generated using Synthetic Fingerprint Generator (SFinGe) [Ma09]. Example images of the FVC06 database are depicted in Figure 2 (a)-(c).
- *MCYT* [Or03]: the fingerprint subcorpus of the MCYT bimodal database contains fingerprint images captured with two different touch-based sensors. Figure 2 (d)-(e) show example fingerprints of this database.
- *PolyU* [LK18]: the Hong Kong Polytechnic University contactless 2D to contact-based 2D fingerprint images database version 1.0 comprises touchless and touch-

(a) FVC06 DB2-A (b) FVC06 DB3-A (c) FVC06 DB4-A (d) MCYT dp (e) MCYT pb

Fig. 2: Example fingerprint images of used subsets of the FVC06 database (a)-(c) and of the MCYT fingerprint subcorpus (d)-(e).

(a) CB-S1 (b) CB-S2 (c) CL-S1 (d) CL-S2

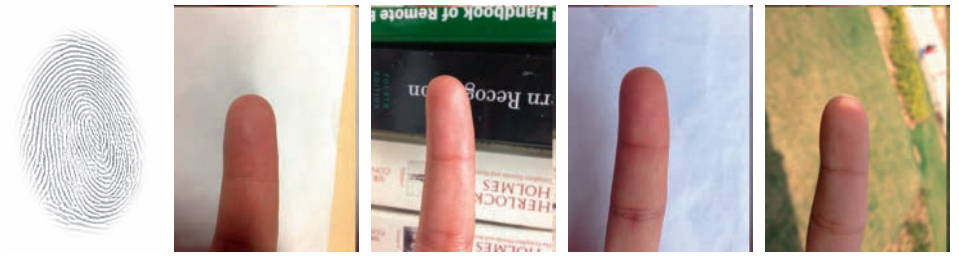
Fig. 3: Example fingerprint images of the subsets of the PolyU database.

based fingerprint images of the same data subjects. In two sessions, touch-based fingerprints were captured with an optical sensor while touchless were acquired using a digital camera with LED illumination. Touchless images, which are provided in pre-segmented form, appear to be captured in a constrained environment. Fingerprint images of this databases are shown in Figure 3.

- *ISPFDv1* [Sa15]: the IIITD SmartPhone Fingerphoto Database v1 consists of touch-based fingerprints captured with an optical sensor as well as touchless fingerprint images collected with a smartphone in four different environmental conditions, including indoor and outdoor images with natural and white background. Figure 4 depicts example images of the ISPFDv1 database. It should be noted that the 250dpi resolution of sensor used to capture the live scan database does not correspond to the NFIQ2.0 target resolution of 500dpi.

3 Evaluation Pipeline

In the proposed evaluation pipeline, touchless fingerprint data is pre-processed, NFIQ2.0 quality scores are estimated, and their predictive power is estimated.



(a) LS (b) WI (c) NI (d) WO (e) NO

Fig. 4: Example fingerprint images of the subsets of the ISPFDv1 database.

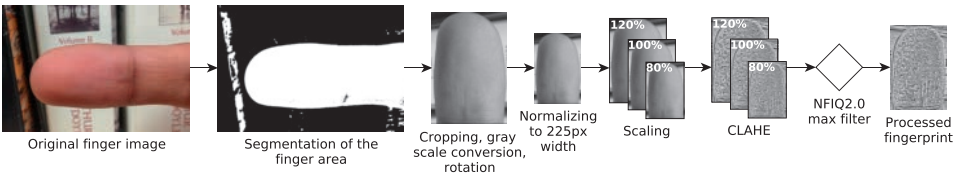


Fig. 5: Proposed touchless fingerprint pre-processing pipeline.

3.1 Touchless Fingerprint Pre-processing

To enable a processing of touchless data with a tool designed for the touch-based domain a pre-processing has to be applied which transforms a touchless fingerprint image to a touch-based equivalent fingerprint image [Sa17,LK18]. The equivalence relates to the resolution of the image respectively to the ridge-to-ridge distance, that can be expected with 10-11 pixels for a 500dpi adult fingerprint. This pre-processing pipeline is necessary since a direct application of a touch-based fingerprint recognition system to touchless fingerprint imagery is not possible [Li13]. Figure 5 depicts the pre-processing pipeline which is used for touchless fingerprint data.

Focusing on the employed touchless databases, two acquisition scenarios can be distinguished: (1) unconstrained acquisition in terms of sensor-to-finger distance, finger rotation, illumination and background properties, which is the case for the ISPFDv1 database, and (2) constrained acquisition, which is the case for the PolyU database. In the latter database, the fingerprint images are already segmented, *c.f.* Figure 3 (c, d). That is, on the PolyU database we skip the segmentation part of the pre-processing pipeline.

To extract the finger area from the background a color-based segmentation method is used [SVC17]. To achieve an accurate segmentation performance, the threshold parameters are adapted for the different environmental situations, *i.e.* subsets of the ISPFDv1 database. In order to present only the fingerprint region to NFIQ2.0 and the feature extractor, a fingertip detection and cropping is performed. Here a brightness-based approach as proposed by Raghavendra *et al.* [RBY13] is used, which searches for the most prominent local minimum on the smoothed gray scale distribution along the horizontal axis. This minimum corresponds to the first finger knuckle. After the cropping step the finger image

only contains the relevant fingertip region. Since the samples of the ISPFDv1 database are represented in a horizontal orientation, samples are rotated by 90 or 270 degree in order to achieve consistency in terms of orientation, *i.e.* upright fingerprint impression. Then the angle between the longitudinal axis of the finger and the horizontal axis is 90 degree.

As can be seen in Table 1, considered touchless datasets consist of color images. Hence, a conversion to gray scale is computed using the very common RGB to gray scale conversion parameters: $Y \leftarrow 0.299R + 0.587G + 0.114B$. Touchless fingerprint samples might be captured at various distances leading to a varying ridge-line frequency. However, the NFIQ2.0 algorithm is designed to achieve optimal results on touch-based fingerprint data captured with a resolution of 500dpi [NI]. For this reason, all touchless samples are normalized to an image width of 225pixel which resembles a ridge-line frequency comparable to that of touch-based fingerprint data captured at a resolution of 500dpi. Due to varying finger sizes and inaccuracies during the finger segmentation a further scaling of $\pm 20\%$ on the normalized image is executed. Assuming that NFIQ2.0 reveals the best scores on 500dpi images we present all three versions of the sample to the NFIQ2.0 method expecting that the one with the highest quality score is the one which is most equivalent to a touch-based capture condition with 500dpi. A max filter is applied and the best quality score represents the final one and the corresponding fingerprint sample is used for further processing.

3.2 Biometric Performance Prediction

For evaluating the predictive power of a quality assessment algorithm for a biometric recognition system Grother and Tabassi [GT07] introduced the ERC. This method evaluates whether a rejection of low quality samples results in a reduce false-non-match error rate (FNMR).Each genuine comparison is associated with a similarity score s_{ii} and two quality scores $q_i^{(1)}$ and $q_i^{(2)}$ in order to aggregate the pair of quality scores from a pair of samples to be compared. As combination function H the min function is chosen:

$$q_i = H\left(q_i^{(1)}, q_i^{(2)}\right) = \min\left(q_i^{(1)}, q_i^{(2)}\right) \tag{1}$$

Then a set $R(u)$ is formed containing the pairwise minima which are less than a fixed threshold of acceptable quality u:

$$R(u) = \left\{i : H\left(q_i^{(1)}, q_i^{(2)}\right) < u\right\} \tag{2}$$

Subsequently, $R(u)$ is used to exclude comparison scores and computing the FNMR on the rest. Starting with the lowest of the pairwise minima, comparisons are excluded up to a threshold t which is obtained by using the empirical cumulative distribution function of the comparison scores, which corresponds to a FNMR of interest denoted by f:

$$t = M^{-1}(1 - f) \tag{3}$$

The ERC is then computed by iteratively excluding a portion of samples and recomputing the FNMR on the remaining comparison scores which are below the threshold:

$$\mathrm{FNMR}(t,u) = \frac{\left|\{s_{ii} : s_{ii} \leq t, i \notin R(u)\}\right|}{\left|\{s_{ii} : s_{ii} \leq \infty\}\right|} \tag{4}$$

Due to the effect that a fraction of low quality samples are excluded in every iteration step the FNMR should decrease constantly if the quality measure is a good predictor for the biometric performance.

In order to compare the different ERCs, the area under each curve minus the area under the optimal curve value is computed and denoted as partial area under curve (PAUC). Here the threshold is set to $x = 0.2$ to consider the most relevant part of the curve only.

4 Experimental Results

In experiments, we first estimate the distributions of NFIQ2.0 scores for touch-based and touchless fingerprint data sets applying the proposed evaluation pipeline. Additionally, the biometric performance is evaluated on the used fingerprint databases employing open-source fingerprint recognition systems. The features (minutiae triplets – 2-D location and angle) are extracted using neural-network based approaches. In particular, the feature extraction method of Tang *et al*. [Ta17] is employed for all databases except for touchless fingerprint images of ISPFDv1 for which the algorithm of Nguyen *et al*. [NCJ18] is applied. The latter feature extractor is designed for more challenging scenarios and hence is more suitable for said image subsets. For both feature extractors pre-trained models are made available by the authors. To compare such templates, a minutiae pairing and scoring algorithm of the sourceAFIS system of Važan [Va19] is used[3]. Moreover, we evaluate the predictive power of NFIQ2.0 regarding biometric performance using the ERC method.

4.1 Sample Quality Estimation

The score distributions of NFIQ2.0 quality scores obtained from the considered databases are plotted in Figure 6. Table 2 lists means and standard deviations of said score distributions together with resulting biometric performance in terms of Equal Error Rates (EERs). EERs are estimated by performing all possible genuine and impostor comparisons. A wide range of quality scores is represented in the NFIQ2.0 score distributions of the FVC06 and MCYT database, *c.f*. Figure 6 (a)-(b). Competitive performance rates are obtained on most subsets of theses databases except for the FVC DB3-A, see Table 2.

By incorporating the proposed pre-processing pipeline for touchless fingerprint imagery, similar NFIQ2.0 quality score distributions can be obtained, *e.g*. for the PolyU database,

[3] The original algorithm uses minutiae quadruplets, *i.e*. additionally considers the minutiae type (*e.g*. ridge ending or bifurcation). Since minutiae triplets are extracted by the used minutiae extractors, the algorithm has been modified to ignore the type information.

Tab. 2: Average NFIQ2.0 scores and biometric performance obtained from the considered databases.

DB	Subset	Preproc.	Avg. NFIQ2.0 score	EER (%)	PAUC
FVC06	DB2-A	–	36.07 (±9.07)	0.15	0.01261
	DB3-A	–	40.92 (±12.85)	6.71	0.00883
	DB4-A	–	27.80 (±12.28)	2.90	0.01261
MCYT	dp	–	37.58 (±15.17)	0.48	0.00868
	pb	–	33.02 (±13.99)	1.35	0.00970
PolyU	CB-S1	–	42.64 (±11.96)	0.67	0.00890
	CB-S2	–	40.97 (±13.14)	1.75	0.00893
	CL-S1	proposed	47.71 (±10.86)	3.91	0.00998
	CL-S2	proposed	47.08 (±13.21)	3.17	0.01106
ISPFDv1	LS	–	58.19 (±7.70)	0.51	0.01275
	NI	proposed	9.62 (±7.65)	34.64	0.01205
	NO	proposed	14.70 (±9.39)	28.12	0.01214
	WI	proposed	16.86 (±7.02)	35.67	0.01465
	WO	proposed	18.60 (±9.77)	25.29	0.01246

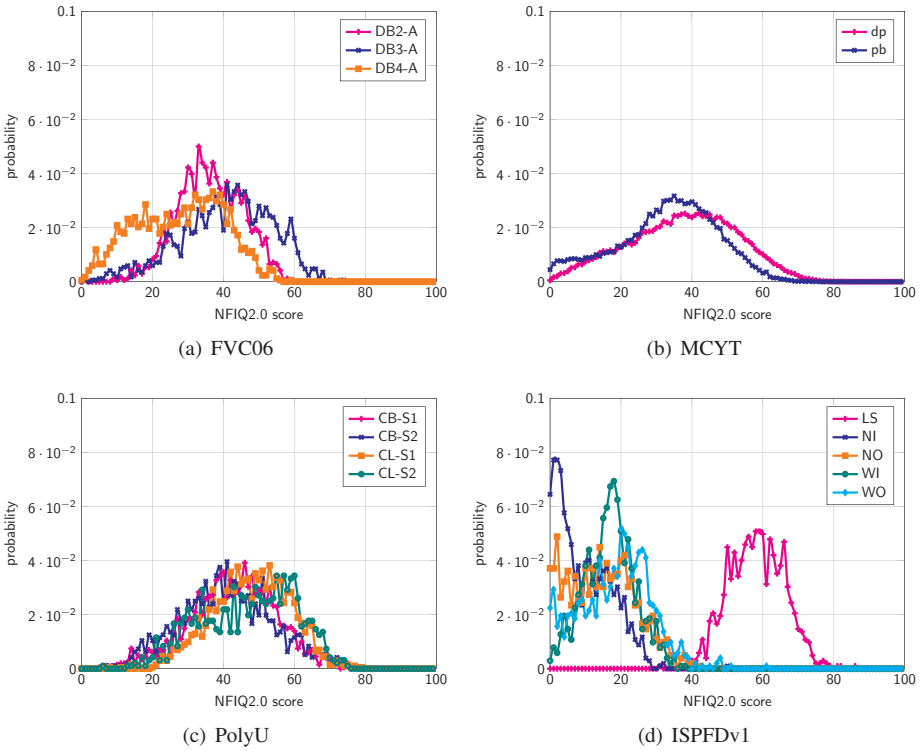
c.f. Figure 6 (c). In contrast, for the ISPFDv1 database two extreme cases can be observed: touch-based fingerprints exhibit very high quality while touchless fingerprint data are of rather very low quality in terms of NFIQ2.0, *c.f.* Figure 6 (d). This can be explained by the fact that the touchless fingerprint data of the ISPFDv1 database was acquired under rather unconstrained conditions, *i.e.* at variable distance, lightning, and focus. This is also reflected by the biometric recognition performance obtained on the subsets of the ISPFDv1 database, see Table 2. In such unconstrained environments dedicated feature extractors are required, as showcased by Sankaran *et al.* [Sa15].

Focusing on the relation of biometric performance and quality score distributions a clear inter-relation between recognition accuracy and quality can be observed from Table 2. However, we also observe that the biometric performance strongly depends on the applied feature extractor. More specifically, lower EERs are obtained for touch-based fingerprint data which has been captured using an optical or capacitive sensor, *e.g.* the MCYT database. In contrast, the fingerprint images of FVC DB3-A and DB4-A, which have been captured with a thermal sensor and generated synthetically, respectively, yield significantly higher EERs albeit exhibiting similar NFIQ2.0 score distributions. This also hold for touchless fingerprint data, as it can be clearly observed from EERs obtained on the PolyU database.

4.2 Biometric Performance Prediction

For the estimation of ERCs a FNMR of 10% is used as starting point for each database as suggested in [OŠB16]. ERCs for the considered databases are depicted in Figure 7. Strongly dropping ERCs indicate high predictive power, *i.e.* the FNMR is effectively reduced by rejecting fingerprint samples which exhibit low quality. Based on the obtained ERCs the following conclusions can be drawn:

- If no significant biometric performance gains are to be expected, the predictive power in terms of ERC is rather low. This corresponds to the cases were either very



(a) FVC06 (b) MCYT

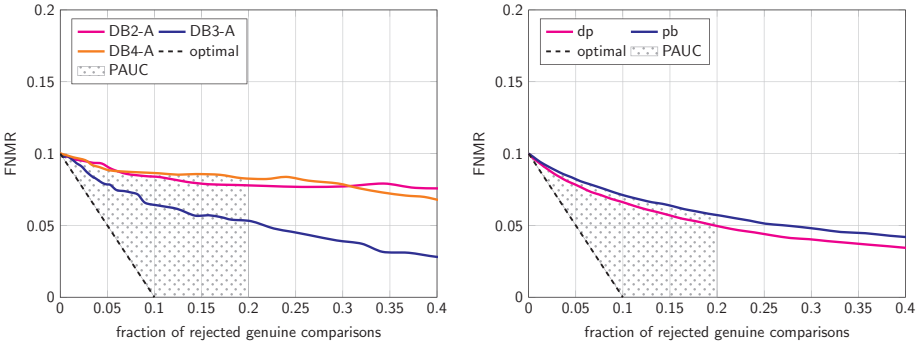
(c) PolyU (d) ISPFDv1

Fig. 6: Probability density functions of NFIQ2.0 scores obtained from the considered databases.

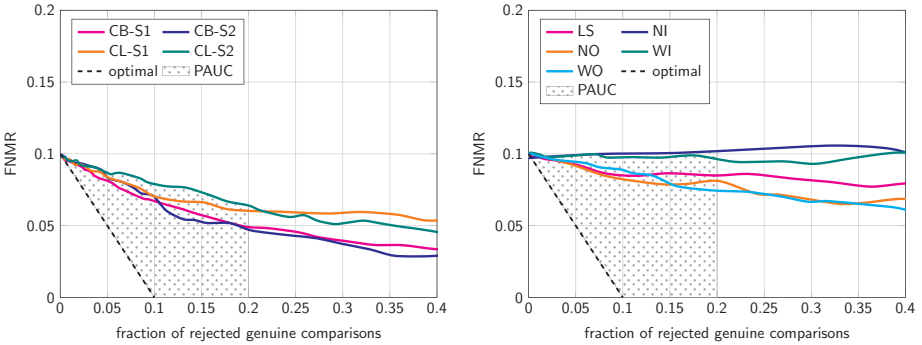
high or very low recognition accuracy is obtained and quality scores are distributed in narrow ranges, *c.f.* ERCs, EERs, and NFIQ2.0 score distributions of FVC06 DB2-A (PAUC: 0.01261) and ISPFDv1 (*e.g.* PAUC NO: 0.01214).

- In case rather low recognition accuracy is obtained or NFIQ2.0 quality score distributions exhibit a wider range, the predictive power in terms of ERC is higher. This can be observed from the ERCs, EERs, and NFIQ2.0 score distributions of FVC06 DB3-A (PAUC: 0.00883), MCYT (*e.g.* PAUC dp: 0.00868), and PolyU (*e.g.* PAUC CB-S2: 0.00893).
- Under the aforementioned condition, the predictive power of NFIQ2.0 for touchless fingerprint data is only slightly inferior compared to that of touch-based fingerprint data. That is, ERCs drop less strongly (*e.g.* PAUC FVC06 DB2-A: 0.01261), *c.f.* ERCs obtained for PolyU (*e.g.* PAUC CB-S2: 0.00893).

Further, it might be concluded that NFIQ2.0 has less predictive power on synthetic data compared to real fingerprint data, *c.f.* ERCs for FVC DB4-A (PAUC: 0.00883).



(a) FVC06 (b) MCYT



(c) PolyU (d) ISPFDv1

Fig. 7: ERCs obtained from the considered databases.

5 Conclusions

This work firstly investigated the applicability and predictive power of NFIQ2.0 for touchless fingerprint data. We conclude that NFIQ2.0 can be a viable tool for quality assessment in touchless fingerprint recognition scenarios in case adequate pre-processing is employed. Finally, it is important to emphasize that more a sophisticated pre-processing might further favour the predictive power of NFIQ2.0 for touchless fingerprint data.

Acknowledgements

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the framework of MEDIAN (FKZ 13N14798)

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

References

[BVS14] Bharadwaj, S.; Vatsa, M.; Singh, R.: Biometric Quality: A Review of Fingerprint, Iris, and Face. EURASIP Journal on Image and Video Processing (JIVP), p. 28, July 2014.

[Ca07] Cappelli, R.; Ferrara, M.; Franco, A.; Maltoni, D.: Fingerprint Verification Competition 2006. Biometric Technology Today, 15(7-8):7–9, 2007.

[GT07] Grother, P.; Tabassi, E.: Performance of Biometric Quality Measures. IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(4):531–543, April 2007.

[La14] Labati, R. D.; Genovese, A.; Piuri, V.; Scotti, F.: Touchless Fingerprint Biometrics: A Survey on 2D and 3D Technologies. Journal of Internet Technology, 15(3):328, 2014.

[Li13] Li, G.; Yang, B.; Olsen, M. A.; Busch, C.: Quality Assessment for Fingerprints Collected by Smartphone Cameras. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 146–153, 2013.

[LK18] Lin, C.; Kumar, A.: Tetrahedron Based Fast 3D Fingerprint Identification Using Colored LEDs Illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12):3022–3033, 2018.

[LPS10] Labati, R. D.; Piuri, V.; Scotti, F.: Neural-based quality measurement of fingerprint images in contactless biometric systems. In: The 2010 International Joint Conference on Neural Networks (IJCNN). pp. 1–8, 2010.

[Ma09] Maltoni, D.; Maio, D.; Jain, A. K.; Prabhakar, S.: Synthetic fingerprint generation. Handbook of fingerprint recognition, pp. 271–302, 2009.

[Ma17] Malhotra, A.; Sankaran, A.; Mittal, A.; Vatsa, M.; Singh, R.: Fingerphoto Authentication Using Smartphone Camera Captured Under Varying Environmental Conditions. In (Marsico, Maria De; Nappi, Michele; Proença, Hugo, eds): Human Recognition in Unconstrained Environments, pp. 119 – 144. Academic Press, 2017.

[NCJ18] Nguyen, D.-L.; Cao, K.; Jain, A. K.: Robust Minutiae Extractor: Integrating Deep Networks and Fingerprint Domain Knowledge. In: The 11th International Conference on Biometrics, 2018. 2018.

[NI] NIST: , NFIQ2.0: NIST Fingerprint Image Quality 2.0. https://github.com/usnistgov/NFIQ2.

[Or03] Ortega-Garcia, J.; Fierrez-Aguilar, J.; Simon, D.; Gonzalez, J.; Faundez-Zanuy, M. et al.: MCYT baseline corpus: a bimodal biometric database. Vision, Image and Signal Processing, IEE Proc. -, 150(6):395–401, December 2003.

[OŠB16] Olsen, M.; Šmida, V.; Busch, C.: Finger image quality assessment features - definitions and evaluation. IET Biometrics, 5(2):47–64, June 2016.

[RBY13] Raghavendra, R.; Busch, C.; Yang, B.: Scaling-robust fingerprint verification with smartphone camera in real-life scenarios. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–8, September 2013.

[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.

[Sa17] Salum, P.; Sandoval, D.; Zaghetto, A.; Macchiavello, B.; Zaghetto, C.: Touchless-to-touch fingerprint systems compatibility method. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3550–3554, 2017.

[SVC17] Sisodia, D. S.; Vandana, T.; Choudhary, M.: A conglomerate technique for finger print recognition using phone camera captured images. In: 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). IEEE, pp. 2740–2746, 2017.

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

[TWW04] Tabassi, E.; Wilson, C.L.; Watson, C.I.: Fingerprint Image Quality. NIST Interagency Report 7151, National Institute of Standards and Technology, August 2004.

[Va19] Važan, R.: , SourceAFIS – Opensource fingerprint matcher. `https://sourceafis.machinezoo.com/`, 2019. Last accessed: August 31, 2020.

[YLB13] Yang, B.; Li, G.; Busch, C.: Qualifying fingerprint samples captured by smartphone cameras. In: 2013 IEEE International Conference on Image Processing. pp. 4161–4165, 2013.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 57

Improved Liveness Detection in Dorsal Hand Vein Videos using Photoplethysmography

Johannes Schuiki[1], Andreas Uhl[1]

Abstract: In this study, a previously published infrared finger vein liveness detection scheme is tested for its applicability on dorsal hand vein videos. A custom database consisting of five different types of presentation attacks recorded with transillumination as well as reflected light illumination is examined. Additionally, two different methods for liveness detection are presented in this work. All methods described employ the concept of generating a signal through the change in average pixel illumination, which is referred to as Photoplethysmography. Feature vectors in order to classify a given video sequence are generated using spectral analysis of the time series. Experimental results show the effectiveness of the proposed methods.

Keywords: near infrared, liveness detection, presentation attack detection, photoplethysmography, dorsal hand vein, video sequence.

1 Introduction

The human body provides a great number of biometric traits which due to their interpersonal uniqueness provide a high level of distinctiveness. Especially identification systems using vascular patterns, that is, structures in blood vessels, have become an important field of research in the last decades. Two important properties of these biometric traits are that they are a) not left behind like fingerprints or DNA and b) usually hidden inside the human body which makes the use of specially designed imaging devices almost inevitable. Common parts of the body for gathering recordings of these vessels are hands (including wrists and fingers) and eyes. This research focuses on information captured from the dorsal view of hand vessels. Such systems make use of the fact that the hemoglobin in the human vessels have a relatively high molar extinction coefficient, which is a measurement for attenuation of electromagnetic waves with respect to a certain wavelength, in the near-infrared (NIR) range. In general one differs between two types of illumination variants (not counting the combination of both techniques into a hybrid version): *Transillumination*, where the hand is placed inbetween the light source and the imaging sensor and *Reflected light*, where the NIR light comes from the same direction as the camera. By doing so, and as the name suggests, the reflected light is then again captured by the imaging sensor. An illustration of these two illumination variants is given in fig. 1 and examples of the resulting images can be seen in fig. 2.
Biometric identification systems in general suffer from what is known as presentation attacks [Ha15]. Here, data is presented to the biometric capturing subsystem with the goal of interfering (e.g. for impersonation) with this system. Countermeasures are therefore

[1] Department of Computer Sciences, University of Salzburg, AUSTRIA, {jschuiki,uhl}@cs.sbg.ac.at

referred to as presentation attack detection (PAD). Usually, PAD methods try to classify the data presented to the capturing subsystem into either attack presentations or bona fide presentations. It has been shown also for the modalities used in this work that presentation attacks exist [PBK16, TM15] and are a potential threat that has to be dealt with. In particular these presentation attacks are conducted by either printing previously acquired bona fide images on paper or utilizing a digital display (e.g. from a smartphone) which is then presented to the imaging sensor. The act of ensuring that the extracted features descent from actual blood vessels is referred to as liveness detection and builds a subset of PAD methods. Related work proposing countermeasures for presentation attacks on NIR vein identification systems can be split into the following categories:

Multimodal approaches, that achieve liveness detection through additional hardware or by looking at more than one trait [Kr18, CT17, CTC18, SBR08]. Other related works provide algorithmic contributions, where authors try to distinguish real and spoof data by software. *Deep learning based* approaches that employ convolutional neural networks (CNNs) build one of those branches. Another idea for PAD is to utilize differences in *quality*, *skin texture* and *spatial frequency components* in still images. A broad overview of the still image approaches (including CNN, quality, texture and frequency) is given in table 14.1 from [Ko20].

A different approach to accomplish PAD is to use *video sequences* that utilize differences of features in adjacent frames. The idea proposed by [Ra15] applies a technique called Eulerian motion magnification on transillumination finger vein videos. This method amplifies minor changes caused by active blood flow. Recently [HU20] analyzed this scheme for a custom dorsal hand vein dataset that was captured with transillumination as well. The following methods build upon a common pre-processing step that allows using the video data for temporal analysis of blood flow in the human hand. With every frame, either on single pixel level, on a region of interest (ROI), or with the image as a whole (for the latter two the average pixel illumination is calculated), a time series is generated. The series should resemble the characteristics of a beating heart, indicating life. This is a form of Photoplethysmography which is explained in section 3. In 2008, [Zh08] used this form of temporal analysis on a ROI of the human hand with reflected light imaging using two different wavelengths, namely 660 and 880nm. The work already contained the observation that one dominant peak in the frequency spectrum around 1.4 hertz would most likely be the heart rate. In [ZH13] the goal was actually to present a new method for palm vein extraction using reflected light imaging, however the procedure includes the estimation of the power spectral density over an 8 minute video sequence. Again, a dominant peak around 1.4 hertz was reported. Another technique for discrimination of bona fide from attack videos was proposed in [Di15]. Here, the average minima and maxima from the time series form features for classification, but it is also shown that this method can be spoofed by a blinking LED that imitates pulse. In [HKL18] again the similarity to blood pressure signals is shown with videos from a custom built NIR transillumination finger vein capturing device, but lacks information on how to use this information for liveness detection.

The work in this paper analyses the applicability of a finger vein liveness detection scheme proposed by Bok et al. [BSL19] on a custom dorsal hand vein dataset explained in section 2. While the reference work uses a database which only consists of sequences captured with transillumination, here we also evaluate the results on reflected light data. Furthermore, this work proposes two additional methods for the construction of a feature vector in order to separate bona fide from presentation attack video sequences. Section 3 explains the existing work and the contributions made by this paper. Section 4 describes the evaluation methods and contains a discussion of the results.

2 Video Database

The imaging installation used to capture the video sequences is similar to the one that was used in [Gr15, HU20]. The imaging sensor is a Canon EOS 5D MarkII DSLR with removed IR blocking filter, that is placed on the top side of a wooden box. Additionally, a 850mn pass-through filter is attached to the scanner. The installation is able to capture data in two illumination variants, namely transillumination & reflected light. In order to accomplish that, a NIR surveillance lamp including 50 940nm NIR LEDs is positioned on the bottom of the capturing device, while 6 950nm LEDs are attached on the top side of the box. Figure 1 illustrates the two modes of operation. The database contains both hands from 13 participants, that were captured with both lighting versions, resulting in 52 genuine video sequences.

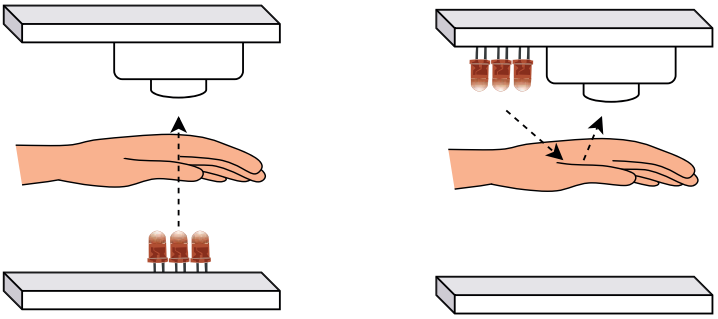


Fig. 1: Modes of operation of the capturing device: Transillumination (left) and Reflected Light (right).

Every video has five presentation attack counterparts, namely a printed frame on cardboard (in this work referred to as *Paper*), the same printed frame but moving back and forth approximately with a pace of 1 hertz to simulate heartbeat artefacts (*Paper Moving*), one frame displayed on a smartphone (*Smartphone*), a frame displayed on a smartphone with a programmed sinusoidal translation (*Smartphone Moving*) and the frame shown on a smartphone with a rhythmic zoom-in-zoom-out effect (*Smartphone Zooming*). Both variants of motion, i.e. translation and scaling, are meant to simulate a heart-beat-like variation of illumination on the dorsum of the hand. For the smartphone attacks a custom android application was created to apply the transformations. Therefore, in total the whole database consists of 312 video sequences, all of resolution 1920x1080. For generation of the attacks, only a region of interest (ROI) was selected. Fig. 2 shows examples of genuine and

attack frames in both illumination variants. The sequences are of varying length. Bona fide videos range from 14.75 to 25.29 seconds, attack sequences from 10.41 to 21.49 seconds. Every sample was captured with a constant frame rate of 29.97 frames per second.

Fig. 2: Example frames from NIR videos; top row: transillumination, bottom row: reflected light; left: bona fide video frame, middle column: paper attack and right: smartphone attack, respectively

3 Remote NIR Photoplethysmography using Spectral Analysis

Plethysmography is the act of measuring changes in volume in different areas of the body. Examples for such measurements are limb circumference, impedance (electric resistance) on certain bodyparts or the amount of air that is being exhaled to conclude lung capacity. One variant of Plethysmography is called Photoplethysmography (PPG) and refers to analysis of optical signals acquired through imaging techniques. PPG makes use of the observation that heart beats generate measureable changes in the human body resulting in a periodical pattern that repeats every cardiac cycle. This non-invasive method is used in devices such as pulse oxymeters or for heart rate estimation in sport watches. In the case of NIR imaging, one utilizes the fact that the oxygen saturation in blood has its peak directly after the heart beat and reaches a minimum shortly before the next impulse. Wei et al. [WHL09] have shown that such blood pressure measurements do not only contain the heart rate, but due to their shape characteristics also include harmonics (i.e. integer multiples of a dominant fundamental frequency) that can be mathematically modelled through exponential decay. This paper exploits this observation for the construction of a feature vector that indicates real blood flow.

A recent publication on presentation attack detection for finger vein image sequences has shown that information from PPG analysis can be used as classification criteria [BSL19]. The proposed approach employs the concept of generating a one dimensional time series by calculation of the average pixel brightness in every frame of the infrared finger vein video. The time series is first zero padded to reach a higher frequency resolution and then transformed into Fourier space using the discrete Fourier transform (DFT). Next, frequency components less than 1.0 and higher than 3.0 hertz are dropped. By taking the magnitude of every frequency component in this range, using a spacing of 0.04 hertz, a 50

dimensional feature vector (FV) is constructed. For classification a support vector machine (SVM) with a radial basis function kernel was used.

3.1 Proposed Methods

Two different ways for FV construction are explained in this section. Both build upon a common basis. An overview is given in figure 3. The methods apply for both illumination variants.

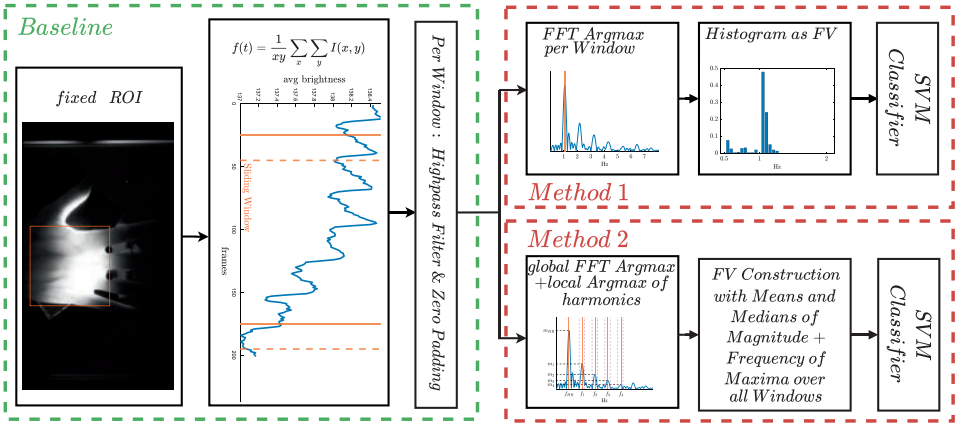


Fig. 3: Block diagram of the proposed methods.

Baseline: Due to the fact that the custom imaging device has two screwed pins where the middle finger has to be placed inbetween, every hand is placed on the same predetermined spot. The pins can also be seen on the subplots of figure 2. Therefore, as a first step, a fixed 600x600 ROI is cropped such that only the dorsal view of the hand (from knuckles to wrist) remains. Similar to the reference method, a time series is built containing values of average pixel illumination (8 bit grey value) from this ROI. In order to center the time series around the x axis, the mean value (DC component) is subtracted. A rectangular window with a size of 150 frames is applied and shifted over the time series with a stepsize of 1 frame. For detrending and removal of other artefacts, an infinite impulse response high-pass filter with a steepness of 0.95 and a cutoff frequency $f_{cut} = 0.5\ Hz$ is applied to every window. As a final step for the baseline, a zero padding of 3000 zeros is attached to the windowed and filtered signal before transformed to frequency space using DFT. The highest frequency and resolution of the spectrum is given by $f_{max} < \frac{29.97}{2} = 14.985\ Hz$ and $\Delta f = \frac{29.97}{150+3000} = 0.0095\ Hz$.

Method 1: For every window, the global *argmax* (i.e. the frequency where the spectrum has its highest peak) is stored in a temporary list. From that list, a histogram is generated with a bin size of 0.05 Hz ranging from 0.5 to 2 Hz covering all windows. Values outside of that range do not contribute to the histogram. The histogram is normalized since video sequences are of varying length and some windows may have its peak outside the range.

The FV for method 1 is given by that normalized histogram and depicts a form of majority voting per window.

Method 2: We define our global $argmax(\mathscr{F})$ as f_{HR} and the $max(\mathscr{F})$ as m_{HR}, where $\mathscr{F}$ is the frequency spectrum ignoring phase information. If f_{HR} is the heart rate, then due to [WHL09] one would expect local maxima at the harmonic frequencies (integer multiples of f_{HR}) with a certain magnitude as well. Therefore, for the first 4 harmonics (i.e. $i * f_{HR}$, $i \in \{2...5\}$) a search window of $\pm\frac{f_{HR}}{5}$ is defined. In addition to f_{HR} and m_{HR}, we temporarily store the local argmax and corresponding maxima as the a quotient w.r.t the global argmax & maxima as f_i and m_i. The final FV is constructed by taking the means and medians (Md) for all stored values over every window, i.e.:

$$\left(\overline{m_{HR}}, \overline{m_1}...\overline{m_4}, Md(m_{HR}), Md(m_1)...Md(m_4), \overline{f_{HR}}, \overline{f_1}...\overline{f_4}, Md(f_{HR}), Md(f_1)...Md(f_4)\right)$$

Exceeding f_{max} with a search window counts as 0 for the entries in question. This case, by construction, can only occur if $f_{HR} * 5 + \frac{f_{HR}}{5} \geq f_{max} \Rightarrow f_{HR} \geq 2.88\,Hz$, which is out of range for a reasonable heart rate.

4 Results

Bok et al. [BSL19] used a SVM classifier with a radial basis kernel function and parameters $C = 10, \gamma = 0.001$. Matlabs *fitcsvm* function was used for classification. Therefore the parameters were set as: 'BoxConstraint' = 10 and 'KernelScale' = $\frac{1}{\sqrt{\gamma}}$. In addition, a simple linear kernel was used for classification. The data under test was split into training and evaluation data using the Leave One Out Cross Validation (LOOCV) principle. Results are reported in compliance with the ISO/IEC 30107-3:2017 standard [IS17], which defines metrics for presentation attack detection such as Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). Table 1 contains the results of both test scenarios. For the proposed methods, the amount of data was given as 26 videos per presentation attack and bona fide respectively. Therefore, the error rate step is $\frac{1}{26} = 3.85\%$. Since for the Bok et al. approach data was split into chunks, between 92 and 109 sequences per category were available for training a classifier. One can see that, with few exceptions, the proposed method 2 is superior to the other two methods under test using the linear kernel classifier. Even with the settings from Bok et al., except for paper based spoofs with reflected light illumination, method 2 yields acceptable results. In comparison to finger vein videos, where the imaging sensor is relatively close to the finger, hand vein imaging offers a lot more possibilities for unwanted errors. As depicted in example time series in figure 4, even slight movements of the hand have noticeable effects in the pixel-averaged time series. That is the reason why using the whole video together with high-pass filtering yields very robust results in comparison to just cutting the sequence into chunks as proposed in [BSL19]. Another measure to circumvent inconsistencies in the time series is given by the windowing. The step size was set to one frame in order to assign more weight on spots where the pulse is clearly visible, which is the case for the major part of the signal. Unfortunately, the proposed approaches are more costly in terms of computational resources which also reflects in the

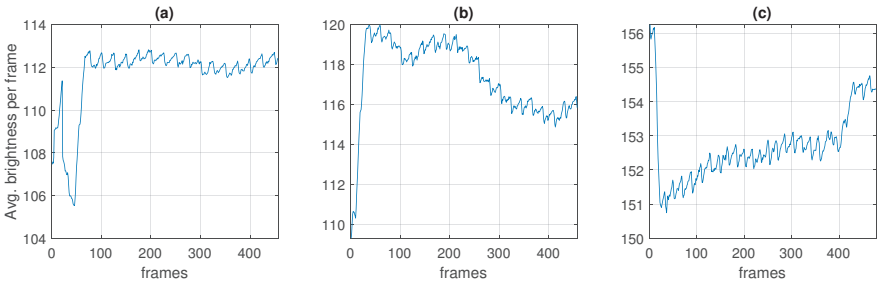


Fig. 4: Types of errors: ascending/descending trend, steps, peaks & fluctuation.

processing time. On average, processing a single video took 32.564 seconds using method 1 and 32.778 seconds using method 2. Compared to the approach from the reference work, where processing all 312 videos (video-slicing included) took 246 milliseconds, this is a downside of the introduced methods. The time consumption is caused by the fact that we apply a fourier transform together with Matlabs highpass filter function on every window with a step size of one frame. The average time consumption per window was 113 milliseconds. As processing hardware, an AMD Ryzen 5 1600X six-core processor was used for all experiments.

5 Conclusion

In this work, the applicability for a finger vein liveness detection scheme was tested on a custom dorsal hand vein dataset in two illumination variants. Two additional classification methods are described. One constitutes a windowed majority voting for the human heart rate by looking for a maximum in the frequency spectrum. The other method exploits the shape characteristic of a typical blood pressure measurement by looking around multiples of the estimated heart rate for other spectral peaks, which has not been used for liveness detection so far. Doing so, the method delivers a certain invariance to the value of the actual pulse and has proven to be a fairly robust approach for both, transillumination and reflected light video sequences. Although the methods are not very efficient in terms of computational cost yet, they form a starting point for a robust unimodal approach for presentation attack detection in dorsal hand vein videos.

6 Acknowledgement

This project was partly funded from the FFG KIRAS project AUTFingerATM under grant No. 864785 and the FWF project ”Advanced Methods and Applications for Fingervein Recognition” under grant No. P 32201-NBL.

	Spoof Method	Bok et al.		Method 1		Method 2	
	RBF: C=10, $\gamma = 0.001$	APCER	BPCER	APCER	BPCER	APCER	BPCER
Transill.	Paper	**10.31**	30.77	11.54	15.38	11.54	**3.85**
	Paper Moving	**4.35**	37.50	26.92	11.54	11.54	**0.00**
	Smartphone	**0.00**	89.42	3.85	7.69	**0.00**	**0.00**
	Smartphone Mov.	77.36	89.42	**0.00**	19.23	19.23	**0.00**
	Smartphone Zoom	59.43	17.31	**3.85**	19.23	11.54	**0.00**
Refl. Light	Paper	**0.00**	80.37	53.85	61.54	7.69	**23.08**
	Paper Moving	49.06	5.61	**23.08**	**3.85**	42.31	7.69
	Smartphone	63.55	6.54	7.69	**3.85**	**3.85**	**3.85**
	Smartphone Mov.	23.08	6.54	3.85	**0.00**	**0.00**	3.85
	Smartphone Zoom	15.24	7.48	3.85	**0.00**	**0.00**	3.85
	linear	APCER	BPCER	APCER	BPCER	APCER	BPCER
Transill.	Paper	11.34	18.27	**7.69**	26.92	11.54	**0.00**
	Paper Moving	**7.61**	16.35	26.92	19.23	11.54	**7.69**
	Smartphone	13.21	54.81	7.69	26.92	**0.00**	**0.00**
	Smartphone Mov.	17.92	57.69	11.54	11.54	**3.85**	**0.00**
	Smartphone Zoom	27.36	40.38	7.69	11.54	**0.00**	**0.00**
Refl. Light	Paper	12.84	71.03	61.54	53.85	**0.00**	**3.85**
	Paper Moving	25.47	**1.87**	15.38	3.85	**3.85**	7.69
	Smartphone	46.73	6.54	7.69	19.23	**3.85**	**3.85**
	Smartphone Mov.	28.85	6.54	3.85	**3.85**	**0.00**	**3.85**
	Smartphone Zoom	18.10	5.61	**3.85**	**0.00**	**3.85**	3.85

Tab. 1: The upper table shows the SVM results with a RBF kernel, BoxConstraint of 10 and a γ value of 0.001 as proposed in [BSL19]; The second table contains results with a simple linear kernel; best results are highlighted **bold**.

References

[BSL19] Bok, Jin; Suh, Kun; Lee, Eui Chul: Detecting Fake Finger-Vein Data Using Remote Photoplethysmography. Electronics, 8:1016, 09 2019.

[CT17] Crisan, Septimiu; Tebrean, Bogdan: Low cost, high quality vein pattern recognition device with liveness Detection. Workflow and implementations. Measurement, 108:207 – 216, 2017.

[CTC18] Crisan, S.; Tebrean, B.; Crisan, T. E.: Multimodal Liveness Detection System for Hand Vein Biometrics. In: 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA). pp. 1–6, 2018.

[Di15] Ding, Henley: Anti-spoofing a Finger Vascular Recognition Device with Pulse Detection. In: 24th Twente Student Conference on IT. University of Twente, Enschede, The Netherlands, 01 2015.

[Gr15] Gruschina, Alexander: VeinPLUS: A Transillumination and Reflection-based Hand Vein Database. In: Proceedings of the 39th annual workshop of the Austrian association for pattern recognition (OAGM15). 2015.

[Ha15] Hadid, A.; Evans, N.; Marcel, S.; Fierrez, J.: Biometrics Systems Under Spoofing Attack: An evaluation methodology and lessons learned. IEEE Signal Processing Magazine, 32(5):20–30, 2015.

[HKL18] Han, Jae Hyun; Kim, Jinman; Lee, Eui Chul: Single-Camera Vision-Based Vein Biometric Authentication and Heart Rate Monitoring via Infrared Imaging Analysis. In (Park, James J.; Loia, Vincenzo; Yi, Gangman; Sung, Yunsick, eds): Advances in Computer Science and Ubiquitous Computing. Springer Singapore, Singapore, pp. 1307–1313, 2018.

[HU20] Herzog, Thomas; Uhl, Andreas: Analysing a Vein liveness detection scheme. In: Proceedings of the 8th International Workshop on Biometrics and Forensics (IWBF'20). Porto, Portugal, pp. 1–6, 2020.

[IS17] ISO: Information technology Biometric presentation attack detection Part 3: Testing and reporting. ISO ISO/IEC 30107-3:2017, International Organization for Standardization, Geneva, Switzerland, 2017.

[Ko20] Kolberg, Jascha; Gomez-Barrero, Marta; Venkatesh, Sushma; Ramachandra, Raghavendra; Busch, Christoph: Presentation Attack Detection for Finger Recognition. In: Handbook of Vascular Biometrics. Springer International Publishing, Cham, pp. 435–463, 2020.

[Kr18] Krishnan, Arya; Thomas, Tony; Nayar, Gayathri R.; Sasilekha Mohan, Sarath: Liveness Detection in Finger Vein Imaging Device Using Plethysmographic Signals. In (Tiwary, Uma Shanker, ed.): Intelligent Human Computer Interaction. Springer International Publishing, Cham, pp. 251–260, 2018.

[PBK16] Patil, I.; Bhilare, S.; Kanhangad, V.: Assessing vulnerability of dorsal hand-vein verification system to spoofing attacks using smartphone camera. In: 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). pp. 1–6, 2016.

[Ra15] Raghavendra, R.; Avinash, M.; Marcel, S.; Busch, C.: Finger vein liveness detection using motion magnification. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–7, 2015.

[SBR08] Shahin, M. K.; Badawi, A. M.; Rasmy, M. E.: A Multimodal Hand Vein, Hand Geometry, and Fingerprint Prototype Design for High Security Biometrics. In: 2008 Cairo International Biomedical Engineering Conference. pp. 1–6, 2008.

[TM15] Tome, P.; Marcel, S.: On the vulnerability of palm vein recognition to spoofing attacks. In: 2015 International Conference on Biometrics (ICB). pp. 319–325, 2015.

[WHL09] Wei, Ching-Chuan; Huang, Chin-Ming; Liao, Yin-Tzu: The exponential decay characteristic of the spectral distribution of blood pressure wave in radial artery. Computers in Biology and Medicine, 39(5):453 – 459, 2009.

[Zh08] Zheng, Jia; Hu, Sijung; Azorin Peris, Vicente; Echiadis, Angelos; Chouliaras, V.A.; Summers, Ron: Remote simultaneous dual wavelength imaging photoplethysmography: A further step towards 3-D mapping of skin blood microcirculation. Proc SPIE, 6850, 03 2008.

[ZH13] Zhang, Huan; Hu, Ding: A Novel Preprocessing Method for Palm Vein. Advanced Materials Research, 658:643–646, 01 2013.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 67

Biometric System for Mobile Validation of ID And Travel Documents

Iurii Medvedev[1], Nuno Gonçalves[2], Leandro Cruz[3]

Abstract: Current trends in security of ID and travel documents require portable and efficient validation applications that rely on biometric recognition. Such tools can allow any authority and citizen to validate documents and authenticate citizens with no need of expensive and sometimes unavailable proprietary devices. In this work, we present a novel, compact and efficient approach of validating ID and travel documents for offline mobile applications. The approach employs the in-house biometric template that is extracted from the original portrait photo (either full frontal or token frontal), and then stored on the ID document with use of a machine readable code (MRC). The ID document can then be validated with a developed application on a mobile device with digital camera. The similarity score is estimated with use of an artificial neural network (ANN). Results show that we achieve validation accuracy up to 99.5% with corresponding false match rate = 0.0047 and false non-match rate = 0.00034.

Keywords: Document security, biometric template, active appearance model, artificial neural network.

1 Introduction

Nowadays, protecting portrait photos on ID and travel documents is of key importance for issuing and legal authorities as face is one of the most largely deployed biometric source [IB08]. That is why the face spoofing attacks widely affect high security field in the companies, government sectors [KSK17]. These attacks usually can be hardly detected by humans as even well trained officers usually perform poorly in matching unfamiliar faces on photos of ID documents, that is why automated systems for efficient document validation are required [SJ19]. Despite all the recent evolution in the facial biometric verification and recognition technologies, when designing security documents and systems, some aspects are relevant. On the one hand, the trend is to allow the validation of documents in totally offline systems, which are designed for scenarios where connectivity may be compromised in terms of availability and security (thus avoiding hacking attacks - such as man-in-the-middle). On the other hand, because of the use of mobile non-proprietary devices, such as smartphones, which are nowadays almost ubiquitous in the hand of authority agents and citizens.

[1] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal, iurii.medvedev@isr.uc.pt
[2] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal; Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal, nunogon@deec.uc.pt, Nuno.MiguelGoncalves@incm.pt
[3] University of Coimbra, Institute of Systems and Robotics - Coimbra, Portugal; Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal, Leandro.Cruz@incm.pt

Consequently, many use cases that rely on facial verification or facial recognition using physical documents are now being designed to allow a fully offline validation with a minutia information extracted from sources of biometric data (such as faces, fingerprints, iris, among others), without storing or accessing databases of face images.

In this paper we present a novel, efficient and compact method for offline mobile applications to secure ID and travel documents with the use of in-house designed facial biometric template and machine readable codes. We are interested in the face photo of documents, either full frontal or token frontal, according to the ISO specifications ISO/IEC 19794-5 [IS11]. Although focused on the face photo, it is worth noting that any source of biometric data (like fingerprint pattern or iris) that can be acquired to perform validation may be employed instead.

The presented approach then solves the document verification task, and does not demand that the biometric samples and features to be stored in any database. This type of validation is sometimes called a match-on-card process.

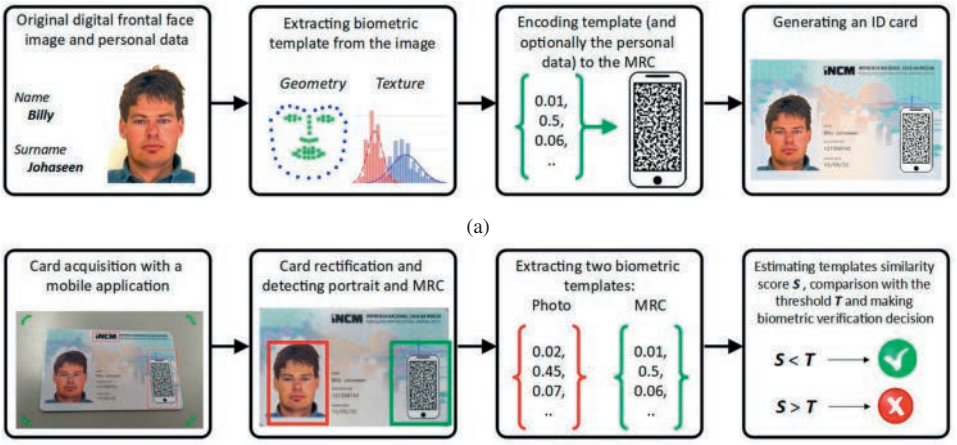


Fig. 1: Pipeline for ID generating(a) and validating(b) in the proposed system.

In our document validation system, we encode the biometric template, that is extracted from a digital frontal face image, within a MRC, and print the resulting code image on the document of an individual (Fig. 1). The approach assumes that the valid MRC for arbitrary biometric sample cannot be recreated by impostors, thus preventing them from issuing the fake identification document that matches different identity. In order to validate the document, the two biometric templates, one from the frontal face photo on its surface and the other from the machine-readable code, are extracted from the testing document to finally be compared to determine if they belong to the same individual. As stated before, this validation shall be performed totally offline. It is important to notice that the biographic data (name, date of birth, etc.) can also be included into the MRC in order to prevent any document ID with original photo and valid MRC to match the data of another individual.

The presented method is related with face recognition problem, however we do not attempt to solve it in its usual formulation. The main goal is to secure a document sample at the moment of its personalization. That is why the problem being solved is the protection of the single document face image itself from biometric impostor attacks like replacing or changing it, into a different identity.

2 Related Work

Recent achievements in different types of face manipulation have made the topic of face forensic recognition to become important for research and investigations [ADB19]. Consequently, document security issues thus encounter new challenges and require efficient methods for face photo security purposes. In [Am19] authors consider a number of different face recognition methods (also including facial landmark-based) in forensics purposes. The paper [DC15] was focused on developing a framework for facial forensics application also related with ID cards fraud. Worth noting that most of the current solutions for document validation are proprietary commercial systems that rely on unpublished algorithms and methods, thus making them difficult to compare with. Consequently, existing benchmarks [GN14] have some submission restrictions.

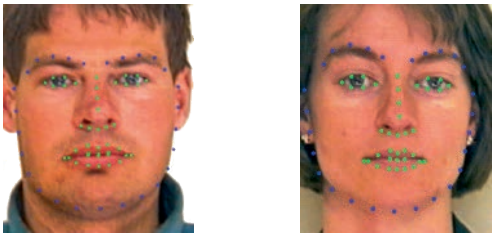
In relation to industry solutions for ID and travel documents, some products have been developed recently. Two approaches are the Digital IPI from Jura [KA04] and both Lasink and DocSeal from IDEMIA [JE19]. They are focused on validation solutions for ubiquitous devices not only to make widely available the access to the authenticity of documents and products, but also to reduce the considerable equipment costs. The main idea of these approaches is to conceal a personalized data within the printed photo that can be further decoded with use of mobile application. Our solution is comparable to these two products, however as they are non-publicly available (both solution and validation dataset), this does not allow one to perform a proper benchmark.

In purposes of face recognition and validation the approaches based on active appearance model (AAM) has been widely used [CET01]. In [ASAAO13], authors use features extracted from AAM of a face and SVM for making a comparison decision. The approach presented in [Ou14] is focused on analysing geometric face distances. The face recognition method in [JP17] is based on face geometric invariances. Modern face recognition approaches usually rely on employing CNN based deep neural networks and use facial landmarks only in alignment [SKP15] or frontalization processes [Ha14]. However, these methods are still computationally heavy especially for application in mobile devices.

3 Facial Biometric Template

Due to limitations of computing power on target platforms, the choice of facial biometic template implementation was made aiming to achieve a trade-off between the calculating complexity and efficiency of its application. We consider the fact that ID and travel documents generally contain frontal face image, where a well-known active appearance model

can be applied (Fig. 2). The model we use contains 68 facial landmarks and denotes the external contour of the face, mouth, eyes, nose, and eyebrows [Re14]. This character of the markup allows one to further determine the various parameters of the face from its image, which can be used for processing by other algorithms.



(a) (b)

Fig. 2: Face image with detected facial landmarks of a man(a) and a woman(b) [Ps].

From the detected set of facial landmarks, we extract values of their coordinates. Nonetheless, those raw data require some normalisation due to the arbitrariness of face location, size and pose on the source image. We solve that problem by performing the following procedure. Firstly, we introduce some predefined set of face feature points. The goal of introducing that supporting set of landmarks is to be the base for aligning other sets of face feature points. If coordinates of two different sets of points are aligned to this supporting set, they will be aligned to each other. As a supporting set in this paper, we choose landmarks detected on artificially generated average face image depicted in Fig. 3a.

In order to align input set of points $\{x_i, y_i\}$ with the supporting set of points, we transform its coordinates to $\{x'_i, y'_i\}$ by rotating, scaling, and shifting it (eq. (1)). The rotation angle α is obtained in order to align input face contour with the horizon (supporting set is already aligned). Although this is a standard procedure, the novelty presented by our work is the scaling, which is performed by the values of face contour perimeter (P_{sup} for supporting set and P for the input set), which is defined by the subset of points with indices 0-26 (blue points on Fig. 2). The shifting vector $S(s_x, s_y)$ is an offset between the average points of supporting and already scaled input sets.

$$\begin{bmatrix} x'_i \\ y'_i \end{bmatrix} = \frac{P_{sup}}{P} * \begin{bmatrix} cos(\alpha) & -sin(\alpha) \\ sin(\alpha) & cos(\alpha) \end{bmatrix} * \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} s_x \\ s_y \end{bmatrix} \tag{1}$$

The set of values that will be a included into a facial biometric template is a result of element-wise subtraction of input and supporting facial landmark coordinates. To avoid depending upon characteristics of images, we divide all the elements of this set to the supporting face contour perimeter P_{sup}. This set of values to be included into the template contains 136 values.

To make the template more robust against biometric distortion attacks (eg. the face image of an impostor can be warped in order to be more geometrically similar to the original identity), texture features are also included into it. In order to extract them, the input face image is aligned with the supporting image, and further segmented in ten characteristic re-

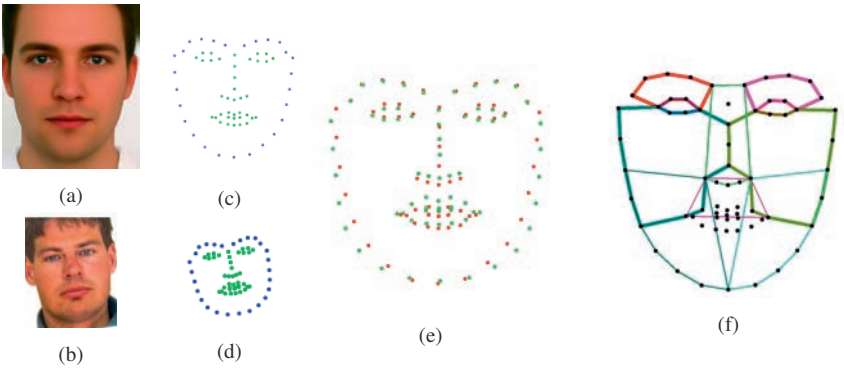


Fig. 3: The process of coordinate recalculating a) supporting face image; b) input face image to be aligned; c) supporting facial landmarks (FL); d) input image FL; e) set of input FL (green) that is aligned with supporting set of FL(red); e) face region contours for extracting HOG features.

gions which are chosen to distinguish face semantic areas (Fig. 3f).Finally for each region features based on histograms of oriented gradients (HOG) are extracted [De11]. This results in 80 features that are concatenated into the resulting biometric template which then contains $D\_size = 216$ elements $\{d_i\}$.

3.1 Template verification

From the ID document that is claimed to be validated, two facial biometric templates are extracted. First $\{d\_test_i\}$ is from the physical face photo that might be forged, and the second $\{d\_orig_i\}$ that is stored on the document, which is assumed to be original. In order to recognize the tested document as genuine, extracted biometric templates are compared between each other. Such straightforward validation can be realised by applying Euclidean distance (eq. (2)). The value of the resulting E in fact indicates the distance score between the templates and can be compared with some threshold in order to trigger the validation decision. Nonetheless, this trivial linear approach does not consider impact weights of different landmarks and seems to be naive and simplistic.

$$E = \sum_{i=1}^{D\_size} |(d\_test_i - d\_orig_i)| = \sum_{i=1}^{D\_size} |d\_sub_i| \tag{2}$$

In order to perform a more robust verification, we solved a binary true-false classification task by designing a multilayer perceptron, where the first layer receives the result of element-wise subtraction of templates. For training that classifier, we employ a classical random sequential back-propagation algorithm [YM98]. The final layer contains one node and returns a response scalar S in the range [0,1].

The input layer of this network receives an element-wise absolute difference of two biometric templates d_sub_i normalised by the coefficient N (see eq. (3)). The purpose of in-

troducing N and limiting the input values to 1 is to fit them within the range of the first layer activation function. In our experiment, the best results are obtained with $N = 0.015$ for geometry based elements and $N = 0.1$ for texture based ones.

$$d\_inp_i = max(1, \frac{|d\_sub_i|}{N}) \tag{3}$$

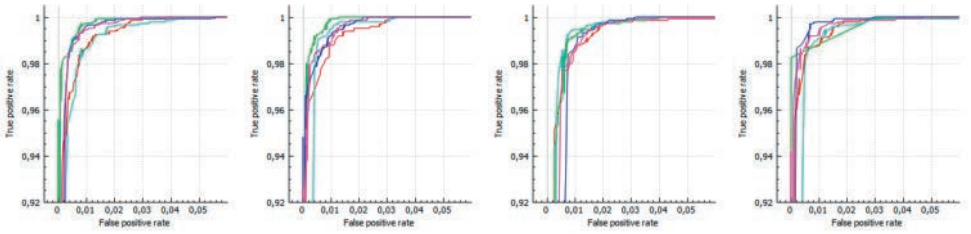
3.2 Classifier Training and Tests

Due to specificity of the verification task, we have prepared an in-house dataset for training, testing and estimating the efficiency of the presented approach. [4] First, a set of frontal face images of 89 individuals was prepared and printed with a size chosen in accordance with [IS19]. In a process of a real document validation with a mobile device, we assume to perform document acquisition with conventional digital camera. In order to follow these conditions properly in training and tests the printing of the dataset is required. These images were used to perform acquisitions similar to Fig. 1a with use of conventional smartphone digital camera (Huawei P20 Pro was used in our tests) and further a perspective transformation based on the structure of the ID document (see Fig. 1b). As a source of frontal face images a dataset from [Ps] was chosen. We have acquired around 4.5 thousand face image samples and in a combination with 89 original digital images, combined around 9 thousand pairs in order to extract a balanced set of $\{d\_sub_i\}$. Each pair of samples contains an original digital image and a rectified capture of the printed image for validating. Pairs with both images belonging to the same identity correspond to the true comparison decision, while the opposite situation imitates a biometric impostor attack of face image replacing. This data was divided into train (70%) and test (30%) parts with different (disjoint) identities in both parts. On this dataset the network learns not only the proper weights for the particular template elements but also learns to avoid noise related with printing, acquisition and rectification inaccuracies.

3.3 Results

For the purpose of choosing a better architecture of the classifier, we have tested a number of different ones. Here we aim to achieve the trade-off between the classifier performance and computing complexity. To evaluate the performance, we build ROC (receiver operating characteristic) curves and estimate their AUC (area under curve) values. (Fig. 4). Architecture with two hidden layers already gives reasonable classification efficiency for our task (Fig. 4d). However the resulting ROC curves are irregular and not smooth, what is not very convenient in practice while tuning to the required optimum between false match and false non-match rates. In our experiments results with less hidden layers were even more impractical. Increasing the number of classifier layers improves results in terms of AUC value (Table 1). At the same time, enlarging the size of the hidden layers does not impact to the performances (Fig. 4c).

[4] https://github.com/visteam-isr-uc/trustfaces-template-verification



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

Fig. 4: ROC curves of ANN classifier with different architectures and different numbers of training epochs. Correspondence between labelled sub-figures and architecture is indicated in Table 1. Different colour of curves indicates different numbers of training epochs.

ANN	**AUC**				
Layers Fig. 4	***red 5ep***	***cyan 10ep***	***green 20ep***	***blue 30ep***	***purple 40ep***
216-300-400-100-1 (a)	0.9996	0.9988	0.99994	0.9992	0.9990
216-300-400-200-100-1 (b)	0.9993	0.9987	**0.99995**	0.9998	0.9998
216-300-500-400-200-1 (c)	0.9985	0.9988	0.9989	0.9979	0.9984
216-300-150-1 (d)	0.9997	0.9988	0.9993	0.9996	0.9998

Tab. 1: AUC of classifiers with different architectures.

For the classifier from the Table 1 with the best value of AUC (architecture: 216-300-400-200-100-1; 20 epochs) we estimated the value of maximum accuracy equal to 0.995 (with the minimum of incorrect classifications). It corresponds to false match rate of 0.0047 and false non-match rate of 0.00034 and is achieved for a threshold value of 0.275.

4 Application of MRC

In order to store the biometric template on the document and make it easily extractable by using a digital camera, we employ machine readable code printed on the document Fig. 1b. Namely we employ the Graphic Code from [CPG18] that can be customised for security purposes. At the stage of creating Graphic Code, several layers of security, robustness and data compressing can be added. The algorithm of creating Graphic Code remains open, even though the Graphic code itself can provide enough computational cost of cryptanalysis by specifying the alphabet that was used in the Graphic Code dictionary, the pattern size, the writing order of cells along the image, the writing order of pixels along the cell, and the dictionary itself. Here we follow the approach of symmetrical encryption where the parameters listed above are the key used both for encryption and decryption, and must be private and secured. Finally, one also can use different methods of cryptography over the data itself, when highly security level is needed. As an example, the message containing facial biometric template can be encrypted to ciphered text what can drastically increase computation complexity of cryptanalysis and security.

Another option is to follow an asymmetric encryption approach. In that case during the decoding process one just needs to prove the document issuer's authority to be sure that the document is not presented by impostor. To achieve that the template data is protected with the digital signature. The issuer authority generates the pair of private and public key. The first one is used to generate the digital signature for the created template that is added to that template to be encoded into the graphic code. With the public key the issuer of the document can be correctly validated. To keep the offline mode of the application that public key must be pre loaded on the device.

4.1 Encoding And Decoding

As a base image for the Graphic Code outline, we use the one depicted on Fig. 5a. Based on a variety of possible unit cells composed of 3×3 pixels, we have defined an alphabet that contains $N = 120$ characters. In order to code the biometric template into the Graphic Code, we transform it into the message in the alphabet space by quantisation process. Each character in the message then correspond to the letter from the alphabet and is replaced by the respective pattern in the dictionary. In addition to the biometric template some information about the individual (ID card number, name) can be also encoded for purposes of automatic document processing. The set of check digits is added to the end of the message. Finally, the remaining cells are replaced using non-dictionary patterns.

(a) (b) (c) (d) (e) (f) (g)

Fig. 5: a) - Graphic Code outline image, b) - example of created Graphic Code acquisition. The process of Graphic Code decoding from physical image: c) - detected Graphic Code image, d) - thresholding and corner detection, e) - base image rectifying, f) - code image extracting and g) - Graphic Code reconstruction.

The decoding of MRC is the inverse to the encoding, and receives the same key parameters that are stored on the device. However, the acquired code image is also needed to be prepossessed and rectified, what is performed with common computer vision algorithms. (Fig. 5c - 5g). During the decoding, the rectified image (Fig. 5f) is overlaid with the grid, then scanned, to find patterns from the dictionary and add corresponding characters to the result message. In that process multiple errors can occur due to various reflection, distortion, MRC surface attrition. That is why after having the full message extracted, its content is validated with the use of check digits. In order to prove the robustness of the decoding we performed extensive tests with the various lighting conditions and applied deformations to the MRC. Most of inaccuracies are mitigated by processing the sequence of multiple camera frames. Only the deformations that significantly damage the MRC unit cells (such as hard scratches) lead to the impossibility of valid decoding.

5 Acknowledgment

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the University of Coimbra for the support of the project TrustFaces.

6 Conclusion

In this paper, we present an efficient and compact method for offline mobile applications to secure ID and travel documents using a facial biometric template and machine readable code. The method demonstrates the high level of efficiency against biometric impostor attacks. This approach solves the frontal face verification problem for purposes of securing ID and travel documents with use of smartphones. Additionally, the presented method of document validation can be expanded for usage with other biometric characteristics (such as fingerprints, iris among others). The practical application does not require sophisticated equipment, thus the approach is also quite cheap in production.

References

[ADB19] Akhtar, Zahid; Dasgupta, Dipankar; Banerjee, Bonny: Face Authenticity: An Overview of Face Manipulation Generation, Detection and Recognition. 05 2019.

[Am19] Amato, G.; Falchi, F.; Gennaro, C.; Massoli, F.; Passalis, N.; Tefas, A.; Trivilini, A.; Vairo, C.: Face Verification and Recognition for Digital Forensics and Information Security. pp. 1–6, 06 2019.

[ASAAO13] Abdulameer, M.; Sheikh Abdullah, S.; Ali Othman, Z.: Face recognition technique based on active appearance model. International Review on Computers and Software, 8:2733–2739, 11 2013.

[CET01] Cootes, T.F.; Edwards, G.J.; Taylor, Christopher: Active Appearance Models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23:681 – 685, 07 2001.

[CPG18] Cruz, L.; Patrão, B.; Gonçalves, N.: Graphic Code: A New Machine Readable Approach. 2018 IEEE International Conference AIVR, pp. 169–172, 2018.

[DC15] Dessimoz, D.; Champod, C.: A dedicated framework for weak biometrics in forensic science for investigation and intelligence purposes: The case of facial information. Security Journal, 29, 12 2015.

[De11] Deniz, O.; Bueno, G.; Salido, J.; De la Torre, F.: Face recognition using Histograms of Oriented Gradients. Pattern Recognition Letters, 32:1598–1603, 09 2011.

[GN14] Grother, Patrick; Ngan, Mei: Face Recognition Vendor Test (FRVT) Performance of Face Identification Algorithms NIST IR 8009. 2014.

[Ha14] Hassner, Tal; Harel, Shai; Paz, Eran; Enbar, Roee: Effective Face Frontalization in Unconstrained Images. 11 2014.

[IB08] IBG: , Biometrics market and industry report 2009-2014. International Biometric Group, Tech.Rep., October 2008.

[IS11] ISO/IEC 19794-5:2011. Information technology — Biometric data interchange formats — Part 5: Face image data. ISO/IEC JTC 1/SC 37 Biometrics, 11 2011.

[IS19] ISO/IEC 39794-5:2019. Information technology — Extensible biometric data interchange formats — Part 5: Face image data. ISO/IEC JTC 1/SC 37 Biometrics, 12 2019.

[JE19] Jones, Robert L.; Eckel, Robert Andrew: , Line segment code for embedding information in an image. U.S. Patent Application No 16/236,068, 2019.

[JP17] Juhong, A.; Pintavirooj, C.: Face Recognition Based on Facial Landmark Detection. 12 2017.

[KA04] Koltai, Ferenc; Adam, Bence: Enhanced optical security by using information carrier digital screening. In (van Renesse, Rudolf L., ed.): Optical Security and Counterfeit Deterrence Techniques V. volume 5310. International Society for Optics and Photonics, SPIE, pp. 160 – 169, 2004.

[KSK17] Kumar, S.; Singh, S.; Kumar, J.: A comparative study on face spoofing attacks. In: 2017 International Conference on Computing, Communication and Automation (ICCCA). pp. 1104–1108, 2017.

[Ou14] Ouarda, W.; Trichili, H.; Alimi, A. M.; Solaiman, B.: Face recognition based on geometric features using Support Vector Machines. In: 2014 6th International Conference SoCPaR. pp. 89–95, 2014.

[Ps] Psychological Image Collection at Stirling (PICS), hhttp://pics.stir.ac.uk/.

[Re14] Ren, S.; Cao, X.; Wei, Y.; Sun, J.: Face Alignment at 3000 FPS via Regressing Local Binary Features. In: 2014 IEEE Conference CVPR. pp. 1685–1692, 2014.

[SJ19] Shi, Yichun; Jain, Anil K.: DocFace+: ID Document to Selfie Matching. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1:56–67, 2019.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A unified embedding for face recognition and clustering. In: 2015 IEEE Conference CVPR. pp. 815–823, 2015.

[YM98] Y.LeCun, L. Bottou, G.B. Orr; Muller, K.-R.: , Efficient backprop, in Neural Networks—Tricks of the Trade. Springer Lecture Notes in Computer Sciences, 1998.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 77

Simulation of Print-Scan Transformations for Face Images based on Conditional Adversarial Networks

Aleksandar Mitkovski[1], Johannes Merkle[2], Christian Rathgeb[2,3], Benjamin Tams[2],
Kevin Bernardo[3], Nathania E. Haryanto[3], Christoph Busch[3]

Abstract: In many countries, printing and scanning of face images is frequently performed as part of the issuance process of electronic travel documents, e.g., ePassports. Image alterations induced by such print-scan transformations may negatively effect the performance of various biometric subsystems, in particular image manipulation detection. Consequently, according training data is needed in order to achieve robustness towards said transformations. However, manual printing and scanning is time-consuming and costly.

In this work, we propose a simulation of print-scan transformations for face images based on a Conditional Generative Adversarial Network (cGAN). To this end, subsets of two public face databases are manually printed and scanned using different printer-scanner combinations. A cGAN is then trained to perform an image-to-image translation which simulates the corresponding print-scan transformations. The goodness of simulation is evaluated with respect to image quality, biometric sample quality and performance, as well as human assessment.

Keywords: Biometrics, face, print-scan transformation, simulation, generative adversarial network.

1 Introduction

Face recognition technologies are frequently utilized for the verification of electronic travel documents, *e.g.*, in automated border crossings. In various countries, the issuance process of electronic travel documents requires applicants to provide face images in digital or analogue form (printed). This has already been identified as security gap since an applicant could manipulate his face image before submitting it to the issuance authority. Possible facial image alterations range from simple retouching [RDB19] to morphing [Sc19], where the latter type of manipulation causes a serious security risk, as shown by Ferrara *et al.* [FFM14]. It was found that human observers achieve only low accuracy in detecting such face image manipulations [RKB17, Rö19]. This necessitates the integration of automated procedures with the aim of reliably detecting face image manipulations.

Recently, different benchmarks [Rö19, Ng20, Ra20] have been conducted to compare the performance of manipulation detection schemes proposed in the scientific literature. The majority of state-of-the-art detection systems relies on machine learning techniques, *e.g.*, deep learning, which usually require a huge amount of training data. In order to ensure

[1] Hochschule Fulda, Fulda, Germany
[2] secunet Security Networks AG, Essen, Germany, {johannes.merkle,christian.rathgeb}@secunet.com
[3] da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Darmstadt, Germany, christoph.busch@h-da.de

high detection performance and generalizability, the used training data should resemble variations present in real-world scenarios. In particular, detection algorithms which are used to analyze face images stored in electronic travel documents are required to exhibit robustness towards print-scan transformations. However, a manual creation of a reasonable amount of printed and scanned face images is time-consuming and costly. Ferrara *et al.* [FFM19] demonstrated that the simulation of print-scan transformations during training can significantly improve the performance of face morphing attack detection on real printed and scanned face images. For this purpose, they employed the scheme of Lin and Chang *et al.* [LC99] which makes use of some mathematical models. Other similar schemes aiming at the simulation of print-scan transformations have been suggested in different application contexts, *e.g.*, [So05, Ei11].

In contrast to published works, we present a GAN-based approach for simulating print-scan transformations. A GAN is a machine learning model in which two neural networks compete with each other to become more accurate in their predictions and to be able to analyze, capture, and resemble the variations within a dataset. A subclass are cGANs which have been found to be suitabe for image-to-image translations [Go14, Is17, Zh17]. Specifically, the term *style transfer* is used to describe the operation of recomposing one image in the style of another (group of) image(s).

In this work, we train a cGAN to perform an image-to-image translation which resembles a print-scan transformation. We obtain face images from two public face database which are printed and scanned employing two printer-scanner combinations. These images are then used together with their original counterparts to train a cGAN to perform an image-to-image print-scan transformation. For each printer-scanner combination, 20 models trained with a different number of epochs are applied to simulate print-scan transformations. Finally, resulting test images are assessed in a comprehensive manner considering many relevant factors such as image quality, biometric sample quality, and human recognition. The obtained results confirm that the proposed cGAN-based approach is capable of realistically simulating print-scan transformations.

This paper is organized as follows: Sect. 2 summarizes the face databases used. The proposed architecture and training of the cGAN is described in detail in Sect. 3. The assessment of the simulated print-scan transformation is presented in Sect. 4. Finally, conclusions are drawn in Sect. 5.

2 Databases

From two public databases, *i.e.*, the FERET [Ph98] and FRGCv2 [Ph05], face images have been manually selected which meet the face image quality standards for electronic travel documents, as specified by the International Civil Aviation Organization (ICAO) [In15]. These images exhibit, among other requirements, full-frontal pose, uniform illumination, good focus, a neutral face expression with open eyes and no visible teeth, and neutral background. For these images, we adjusted the alignment of the face by suitable scaling, rotation and padding/cropping to ensure that the ICAO requirements with respect to the

eyes' positions are met. Overall, 529 and 984 face images have been selected from the FERET and FRGCv2 databases, respectively. Subsequently, all face images are printed and scanned using two printer-scanner combinations listed in Table 1. The resulting image sets are referred to as IS1 and IS2. That is, both image sets contain all 1,513 face images selected from the FERET and FRGCv2 databases, but have been printed and scanned using different devices. Example face images of the original database, IS1, and IS2 are shown in Fig. 1.

Tab. 1: Printer-scanner combinations used for database creation.

Image set	Printer	Scanner	Properties
IS1	Fujifilm Frontier 5700R Minlab	Epson DS-50000	300 dpi, 24-bit RGB, print on matte photo paper
IS2	Developed by professional photo studio	Canon Imagerunner Advance 4535i	600 dpi, 24-bit RGB, print on glossy paper

(a) original (b) IS1 (c) IS2

Fig. 1: Example images of used databases. Best viewed in electronic format (zoomed in).

The printing and scanning process was conducted in a semi-automatic manner. For this purpose, a software tool was implemented which enables the arrangement of images in the size of passport face images according to ICAO [In15] on A4 paper sheets, 20 images per page. Additionally, markers were included in the top left and the right bottom corners of each face image to facilitate a subsequent segmentation. Further, the filenames of images where included as QR-codes. The resulting sheets were then printed on photo paper and scanned, *cf.* Table 1. Finally, the face images were automatically extracted from the scanned sheets and the corresponding filenames were assigned. Fig. 2 shows examples of face images after printing and scanning.

3 Proposed approach

A GAN architecture [Go14] consists of a generator model G for outputting synthetic images according to a given distribution, and a discriminator model D that classifies images as real (from the dataset) or fake (generated). The discriminator model is updated directly,

Fig. 2: Example images of after printing and scanning with segmentation makers and filenames in form of QR-codes. Best viewed in electronic format (zoomed in).

whereas the generator model is updated via the discriminator model. As such, the two models are trained simultaneously in an adversarial process where the generator seeks to fool the discriminator and the discriminator seeks to better identify the generated images.

In a cGAN, G generates the images not just from internal noise z (as in a tradional GAN) but also based on an input image x; the conditional distribution of the output image $G(x,z)$ given the input x is supposed to resemble that of real image translations (x,y). The discriminator is provided both with a source image x and a target image, and must determine whether the target is a real image y or an output $G(x,z)$ of the generator. An example is the well-known Pix2Pix framework of Isola *et al.* [Is17]. The generator is trained via adversarial loss,

$$\mathscr{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x,G(x,z))], \tag{1}$$

which encourages the generator to generate plausible images in the target domain. The generator is also updated via L1 loss measured between the generated image and the expected output image,

$$\mathscr{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x,z)||_1]. \tag{2}$$

This additional loss helps the generator model to create translations nearer to the ground-truth, resulting in,

$$G^* = \arg\min_G \max_D \mathscr{L}_{cGAN}(G,D) + \lambda \mathscr{L}_{L1}(G). \tag{3}$$

The Pix2Pix cGAN has been demonstrated on a range of image-to-image translation tasks such as converting maps to satellite photographs, black and white photographs to color, and sketches of products to product photographs [Is17].

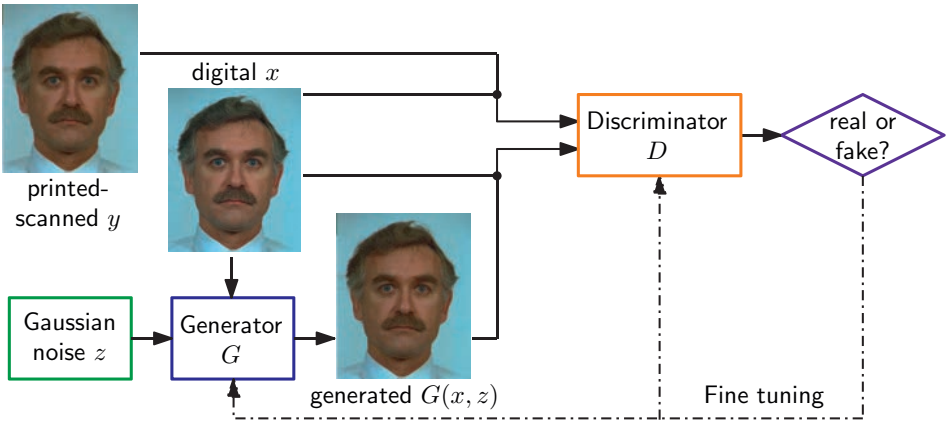


Fig. 3: Concept of the proposed cGAN-based print-scan transformation.

Our approach builds on the Pix2Pix framework [Is17] and is depicted in Fig. 3. Alternatively, a Cycle-Consistent Adversarial Network (CycleGAN) [Zh17] could be employed which allows for the training of an unpaired image-to-image translation. However, since a sufficient number of images is available in digital and printed and scanned form, the Pix2Pix framework turns out to be suitable.

As suggested in [Is17], an image classifier named PatchGAN is used instead of a traditional discriminator. While a traditional discriminator maps the complete image to a single scalar which expresses the probability whether the image is real or generated, the PatchGAN splits the image into small local patches. The L1 norm in the loss function already encourage the generator to correctly represent the coarse structures of the target image. Therefore, the discriminator must only assess if the generator's outputs resemble the fine structures in a realistic way. To this end, the discriminator only requires small patches of the image as input. By restricting the input of the discriminator to small images patches, the size of the discriminator, and number of its parameters can be greatly reduced. In our case, the coarse structures, for which correct representation is encouraged by the L1 norm, are the features of the subject depicted in the input image which are still visible in the print-scan transformed image, and the fine structures investigated by the discriminator are the artefacts induced by the print-scan transformation. For computing the loss function, the images are divided in $N \times N$ patches, and after passing the corresponding pairs of patches to the discriminator the resulting outputs are averaged to estimate whether the image is real or generated.

We used a PatchGAN with fixed kernel size of 4×4, a fixed stride of 2×2, and 70×70 patches as input. During training, horizontal mirroring was applied for data augmentation. This resulted in approximately 3,000 face images per image set. The training on each image set has been performed with 2,400 randomly chosen face image pairs. Training has been conducted separately for each image set with up to 200 epochs each and with a batch size of 1, as suggested in [Is17]. Each epoch uses the entire training set. One

data batch passes the neural network 480,000 times. For GANs, it is often difficult to find the time when training should be stopped. Therefore, we save and evaluate the model after every 10 epochs, resulting in 20 models for each image set. Subsequently, print-scan transformations have been performed on 100 randomly chosen face images of the remaining ones. Note that the number of test images was restricted by the database size as well as time constraints. Examples of simulated print-scan transformations are depicted in Fig. 4.

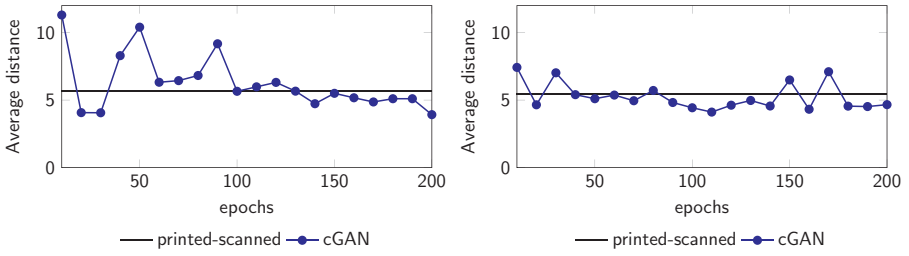
(a) digital (b) printed-scanned (c) generated

Fig. 4: Example images of IS1 (top row) and IS2 (bottom row). Models trained with 180 epochs were used to generate images in the rightmost column. Best viewed in electronic format (zoomed in).

4 Assessment

Firstly, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [MMB12] is used to determine an appropriate number of epochs used during training. BRISQUE calculates a no-reference image quality score which is sensitive to various distortions, *e.g.* blur. Three scores, s_d, s_p, and s_g, are estimated for a digital face image and its printed-scanned as well as generated counterpart, respectively. For such triples of face images we then estimate the distances of scores of digital images to printed-scanned, *i.e.*, $|s_d - s_p|$, and to generated ones, *i.e.*, $|s_d - s_g|$. Eventually, the average distance μ is reported. Thereby, we

measure whether the generated images resemble the effects of a real print-scan transformation. Note that BRISQUE scores of s_d are generally smaller than those of s_p or s_g. Further, note that since the generator models a probabilistic function that describes the image modifications induced by the print-scan transformation, its output cannot perfectly match the target images. The results are plotted in Fig. 5. In partcular on IS1, higher variations can be observed for training with less 100 epochs. For the subsequent experiments (including human assessment) we only consider cGANs which have been trained with 180 or 190 epochs as these configurations yield good results on both image sets.



(a) IS1 (b) IS2

Fig. 5: Average distance of BRISQUE scores across training epochs on both image sets.

In the second experiment, the effects of printing and scanning on face recognition are analyzed, in particular sample quality assessment and performance. To this end, the FaceQNet algorithm [He19] and the ArcFace system [De19] are used for quality estimation and comparison, respectively. The impact on sample quality is computed using the scores of FaceQNet and ArcFace in the same manner as for BRISQUE scores before. In the case of recognition, unconstrained probe face images have been additionally chosen from the FERET and FRGCv2 face databases. Lastly, the Root-Mean-Square Error (RMSE) between the digital reference image and the printed-scanned as well as generated image is estimated for each triple of face images. To do so, image pairs are firstly aligned using the Scale-Invariant Feature Transform (SIFT)-based Fiji tool [Lo99, Sc12] and the RMSE is estimated for each color channel. As final score, the average RMSE over all color channels is calculated. All assessment algorithms process the cropped face regions since these facial image parts are most relevant. It is important to note that the considered assessment algorithms produce scores in different ranges. Obtained results including average distances μ and the standard deviations σ for the real printed-scanned images and the cGAN-based approach trained with different numbers of epochs (ep.) are summarized in Table 2.

Tab. 2: Obtained results for different quality assessment algorithms on both image sets.

Algorithm	**FaceQNet** ($\times 10^2$)				**ArcFace** ($\times 10^2$)				**BRISQUE**				**RMSE**			
Image set	**IS1**		**IS2**		**IS1**		**IS2**		**IS1**		**IS2**		**IS1**		**IS2**	
Score	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Printed-scanned	1.01	0.796	0.892	0.634	1.96	1.52	0.524	0.421	5.64	4.1	5.46	3.73	12.4	2.55	21.0	2.75
cGAN (180 ep.)	0.995	0.746	0.946	0.671	2.68	2.12	0.471	0.348	5.1	3.92	4.55	3.57	12.9	2.95	22.4	2.47
cGAN (190 ep.)	1.12	0.78	1.02	0.654	2.87	1.95	0.512	0.382	5.48	3.92	4.52	3.4	12.9	2.97	22.0	2.75

It can be seen that for the considered assessment algorithms, the generated images yield effects similar to those of the real printed-scanned images. With only a very few outliers,

e.g., effects on ArcFace scores in IS1, this is true for both used printer-scanner combinations. Note that the proposed approach also resemble image set specific variations which can be observed from ArcFace and RMSE scores.

In addition to the above evaluation, a human assessment has been conducted in a second experiment. For this purpose, 35 experts form secunet Security Networks AG were asked to rate images with one to three stars depending on how closely these resemble real printed and scanned face images. After an explanatory introduction to the experiment, they were presented with 20 triples of face images, *i.e.* a real printed-scanned image, a cGAN-based generated image, and an image to which a style transfer based on DeepArt[3] has been applied, see Fig. 6. The latter images were created by transforming the digital images providing their printed-scanned counterpart image as desired style image to the web-based DeepArt style transfer application. Triples of face images were presented to the participants in a randomized order. Obtained results in terms of average rating are shown in Fig. 7.

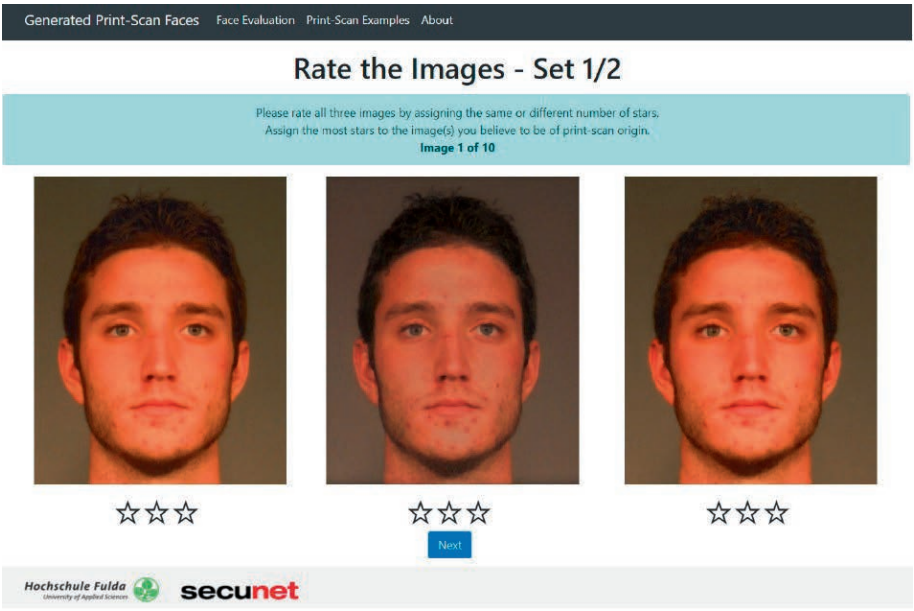


Fig. 6: Screenshot of the rating application in the human assessment experiment.

In can be observed that the proposed cGAN-based approach outperforms the DeepArt style transfer in terms of visual perception. The ratings obtained by the cGAN-based simulation of print-scan transformations are close to those given for real printed-scanned face images.

5 Conclusions

The aim of this work was to generate cGAN-based images that are virtually indistinguishable from real printed and scanned images. In a comprehensive assessment it has been

[3] DeepArt: `https://deepart.io/`

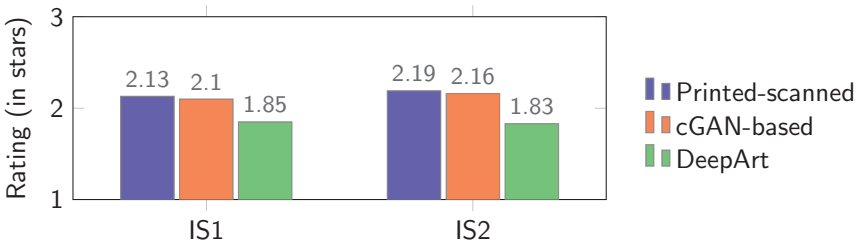


Fig. 7: Obtained results for the human assessment experiment.

shown that the presented approach is capable of simulating the effects of real printed and scanned face images for different printer-scanner combinations. That is, our cGAN-based simulation of print-scan transformations can be used to automatically generate training data as input for face image manipulation detection systems which is subject to future work.

Acknowledgements

This research work has been partly funded by the Federal Office of Information Security (BSI) through the FACETRUST Project, the German Federal Ministry of Education and Research and the Hessen StateMinistry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

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

[Ei11] Eid, A. H.; Ahmed, M. N.; Cooper, B. E.; Rippetoe, E. E.: Characterization of Electrophotographic Print Artifacts: Banding, Jitter, and Ghosting. IEEE Trans. on Image Processing, 2011.

[FFM14] Ferrara, M.; Franco, A.; Maltoni, D.: The magic passport. In: Int'l Joint Conf. on Biometrics (IJCB). 2014.

[FFM19] Ferrara, M.; Franco, A.; Maltoni, D.: Face morphing detection in the presence of printing/scanning and heterogeneous image sources. CoRR, abs/1901.08811, 2019.

[Go14] Goodfellow, I.; Mirza, J. Pouget-Abadie M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y.: Generative Adversarial Nets. In: Advances in Neural Information Processing Systems (NIPS). 2014.

[He19] Hernandez-Ortega, J.; Galbally, J.; Fiérrez, J.; Haraksim, R.; Beslay, L.: FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In: Int'l Conf. on Biometrics (ICB). 2019.

[In15] International Civil Aviation Organization: ICAO Doc 9303, Machine Readable Travel Documents – Part 9: Deployment of Biometric Identification and Electronic Storage of Data in MRTDs (7th edition). Technical report, ICAO, 2015.

[Is17] Isola, P.; Zhu, J.; Zhou, T.; Efros, A. A.: Image-to-Image Translation with Conditional Adversarial Networks. In: Conf. on Computer Vision and Pattern Recognition (CVPR). 2017.

[LC99] Lin, C.; Chang, S.: Distortion Modeling and Invariant Extraction for Digital Image Print-and-Scan Process. In: Int'l Symposium on Multimedia Information Processing (ISMIP). 1999.

[Lo99] Lowe, David G.: Object Recognition from Local Scale-Invariant Features. In: Int'l Conf. on Computer Vision (ICCV). 1999.

[MMB12] Mittal, A.; Moorthy, A. K.; Bovik, A. C.: No-Reference Image Quality Assessment in the Spatial Domain. IEEE Trans. on Image Processing, 2012.

[Ng20] Ngan, M.; Grother, P.; Hanaoka, K.; Kuo, J.: Face Recognition Vendor Test (FRVT) Part 4: MORPH Performance of Automated Face Morph Detection. Technical Report NISTIR 8292, National Institute of Technology (NIST), 2020.

[Ph98] Phillips, P. J.; Wechsler, H.; Huang, J.; Rauss, P. J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 1998.

[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: Conf. on Computer Vision and Pattern Recognition (CVPR). 2005.

[Rö19] Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. In: Int'l Conf. of Computer Vision (ICCV). 2019.

[Ra20] Raja, K.; Ferrara, M.; Franco, A. et al.: , Morphing Attack Detection – Database, Evaluation Platform and Benchmarking, 2020. arXiv 2006.06458.

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

[RKB17] Robertson, D. J.; Kramer, R. S. S.; Burton, A. M.: Fraudulent ID using face morphs: Experiments on human and automatic recognition. PLOS ONE, 2017.

[Sc12] Schindelin, J.; Arganda-Carreras, I.; Frise, E. et al.: Fiji: an open-source platform for biological-image analysis. Nat Meth, 2012.

[Sc19] Scherhag, U.; Rathgeb, C.; Merkle, J.; Breithaupt, R.; Busch, C.: Face Recognition Systems under Morphing Attacks: A Survey. IEEE Access, 2019.

[So05] Solanki, K.; Madhow, U.; Manjunath, B. S.; Chandrasekaran, S.: Modeling the print-scan process for resilient data hiding. In: Security, Steganography, and Watermarking of Multimedia Contents VII. 2005.

[Zh17] Zhu, J.; Park, T.; Isola, P.; Efros, A. A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: Int'l Conf. on Computer Vision (ICCV). 2017.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 87

Unit-Selection Based Facial Video Manipulation Detection

Thomas Nielsen[1], Ali Khodabakhsh[2], Christoph Busch[3]

Abstract: Advancements in video synthesis technology have caused major concerns over the authenticity of audio-visual content. A video manipulation method that is often overlooked is inter-frame forgery, in which segments (or units) of an original video are reordered and rejoined while cut-points are covered with transition effects. Subjective tests have shown the susceptibility of viewers in mistaking such content as authentic. In order to support research on the detection of such manipulations, we introduce a large-scale dataset of 1000 morph-cut videos that were generated by automation of the popular video editing software Adobe Premiere Pro. Furthermore, we propose a novel differential detection pipeline and achieve an outstanding frame-level detection accuracy of 95%.

Keywords: Morph-cut, Video Manipulation, Interframe Forgery, Dataset, Video Manipulation Detection, Video Authenticity.

1 Introduction

Following the evolution of artificial intelligence and the rapid increase in the computational capacity of computers in recent decades, many novel video manipulation techniques have been introduced and became feasible. Despite the original intention of the developers of these techniques, many of them have the potential of being misused by malicious actors to spread disinformation for political and financial aims. Following the significant media attention to this problem after the introduction of Deepfakes, many research groups attempt to address the vulnerability [Ve20]. However, among video manipulation techniques, vulnerability to unit-selection based methods have been overlooked. Unlike Deepfakes and similar generation methods for which synthesis still requires a significant amount of expert knowledge and computational capacity, unit-selection based video manipulation can be flexibly done by commercial software such as Adobe Premiere Pro through their easy to use graphical user interface. Furthermore, subjective tests have shown unit-selection based manipulations to be more difficult to detect for humans than intra-frame manipulations [KRB19]. The use of seamless cut-point transitions is commonplace in media for shortening and summarizing the highlights of videos and they go unnoticed more often than not[4].

Due to the less computational cost and the higher video-realism of unit-selection based generation methods, these methods have been explored for synthesis early-on for appli-

[1] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 324, Kgs. Lyngby, Denmark, s144458@win.dtu.dk
[2] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Teknologiveien 22, Gjøvik, Norway, ali.khodabakhsh@ntnu.no
[3] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Teknologiveien 22, Gjøvik, Norway, christoph.busch@ntnu.no
[4] https://metro.co.uk/2018/12/13/viewers-baffled-child-appears-teleport-tv-interview-8244024/

cations like audio-visual synthesis and video dubbing [MV15]. Even though concatenative generation methods require long videos with constrained recording conditions to be seamless, thanks to searchable public archives of videos, there exists enough footage from interviews on celebrities and political figures for these methods to be feasible. The first automatic technique for face-animation was proposed by Bregler et al. in 1997 [BCS97]. They create a database of visemes[5] from existing footage and, given an input text, they retrieve the visemes and concatenate them using morphing to synthesize a new sentence. More recently, Berthouzoz et al. [BLA12] introduced an editing tool to place visible cuts and seamless transitions in interview videos based on text transcript, which was further developed into the morph-cut transition in Adobe Premiere Pro[6] as a replacement for B-roll[7] and jump-cut transitions[8] for video summarization. Mattheyses and Verhelst [MV15] and Johnston and Elyan [JE19] provide an overview of existing unit-selection based manipulation methods. Among the existing datasets, the biggest that includes inter-frame forgery is VTD 2016 [ASAS16] which is comprised of 33 videos, 6 of which contain inter-frame forgery. Johnston and Elyan [JE19] provide a review of existing video tampering datasets.

In the context of facial video manipulation, a substantial amount of research is oriented towards intra-frame facial video manipulation detection [Ve20]. However, there exists a gap in knowledge with regards to detection of unit-selection based facial video manipulation, and to the best of our knowledge, there are no dataset and no proposed detection method that explicitly address this vulnerability. Nonetheless, Among the proposed methods for the detection of intra-frame manipulations, some utilize inter-frame information for detection to a limited extent. The authors in [GD18] and [Sa19] exploit the inter-frame dependencies to detect frame-by-frame manipulations via a convolutional long short-term memory (LSTM) network and a recurrent neural network respectively. Amerini et al. [Am19] use estimation of the optical flow field as input to a convolutional neural network (CNN) for the detection of inter-frame inconsistencies.

To reduce the visibility of concatenation points in inter-frame forgery, simple gradual transitions such as interpolation, warping, and morphing, as well as more advanced methods such as face-specific warping [Da11] and intermediate frame mining [BLA12] can be used. Examples of advanced transitions that are already available in video editing software are Adobe Premiere Pro Morph-cut (Figure 1) and Avid[9] Fluid Morph. Despite the core algorithms of these transitions being trade secrets, the name of these transitions implies the use of morphing in some form. Consequently, single-image face morphing detection algorithms that are developed in the context of biometric presentation attack detection become relevant for detection. Scherhag et al. [Sc19] provide a recent survey of existing morphing attack detection methods. Asaad and Jassim [AJ17] used the responses of uniform local binary pattern (LBP) extractors on the image to build a Vietoris-Rips complex for detection.

[5] Visemes denote the shape of the mouth when pronouncing specific phonemes. Visemes and phonemes do not share a one-to-one correspondence.
[6] https://www.adobe.com/products/premiere.html
[7] In B-roll transition, a supplemental footage is intercut with the main shot to cover the cuts.
[8] In jump-cut transition, the cut is kept as it is, causing an abrupt jump in the resulting footage.
[9] https://www.avid.com/

Wandzik et al. [WKG18] use high-level features of pretrained face recognition networks as input for a linear SVM classifier.

Fig. 1: An example of a morph-cut transition.

Another set of relevant detection methods can be adopted from general-purpose inter-frame forgery detection, namely frame-insertion and frame-deletion detection methods. Siatara and Mehtre [SM16] provide an overview of the existing inter-frame forgery detection methods. Notably, Chao et al. [CJS12] detect manipulated videos by using the consistency in the total optical flow values in the X and Y directions. More recently, Bakas and Naskar [BN18] used 3D convolutional neural networks with a special difference layer to detect out of place frames in the video sequence.

In this work, we introduce a large-scale dataset of videos containing morph-cut transitions based on videos collected from the wild.[10] To the best of our knowledge, the Morph Cut dataset is the first of its kind and enables the training of deep learning solutions for the detection task. Furthermore, we introduce a robust neural detection pipeline, capable of detecting the morph-cut position at the frame level in a video. The rest of this article is organized as follows: The dataset and the proposed detector are introduced in Section 2. The experiment setup is explained in Section 3 and the results are discussed in Section 4. Finally, the paper is concluded in Section 5.

2 Methodology

Due to the lack of datasets containing a sufficiently large number of unit-selection based manipulation in the literature, we decided to generate a dataset and provide it publicly to stimulate further research in inter-frame forgery detection. In this section, we summarize the construction process of the new Morph Cut dataset along with the description of our proposed method for detecting the inter-frame forgeries.

2.1 Morph Cut Dataset

The development of deep learning-based detectors requires large-scale datasets. Consequently, as the manual generation of datasets of such scale is impractical, the generation process needs to be automated. Adobe Premiere Pro is a well-known popular video editing application that features a seamless morph-cut transition for cut-point concatenation. Furthermore, Adobe Systems provide the scripting language named Extendscript which can

[10] The instructions on how to download the Morph Cut dataset are available at `http://ali.khodabakhsh.org/research/morphcut/`

be used for automation of repetitive tasks in video editing. As such, Adobe Premiere Pro morph-cut transition is the perfect candidate to be used for the generation of the dataset. To achieve a seamless transition, the frames before and after transition need to be similar with regards to the background as well as the general body posture.

To ensure the quality of the generated data, we relied on a much larger video dataset consisting of interview videos as the basis for video selection. Thereafter, based on the movements of face bounding-box after face detection in the videos and the structural similarity of the frames to one another, the videos were ranked and the most suitable videos were selected for the application of morph-cut. Subsequently, the transition is applied to the videos at random points during the interview and the resulting manipulated videos were manually investigated for videos with visible artifacts to be discarded.

2.2 Morph-cut Detection

The unit-selection based video synthesis requires smooth transitions at the cut-points to cover the abrupt changes between the frame before and after. As such, it is safe to assume the existence of frame interpolation during the transition in one form or another. During frame interpolation, the content of the new frame in-between is generated based on the information available in the frame before and after. In contrast, pristine frames contain a natural variability that is not completely explainable based on the information in the frame before and after. Let us consider the frame in the middle to be consisting of two factors, p for the redundant information that is inferable from the frame before and after, and u for the unpredictable natural variability. A good frame interpolation would be able to infer p accurately, however, inference of u is an ill-defined problem. If during the design and training of a frame interpolation method, no mechanism is considered for ignoring u, the objective function would force the interpolation method to generate an average u which minimizes the penalty, yet never occurs in the pristine data. This phenomenon often results in synthetic samples described as over-smooth.

Considering any two frame interpolation methods with the aforementioned characteristics, we hypothesize that the predicted intermediate frames would show more similarity to each other than to the pristine data. The rationale behind this is that the p factor would exist in both pristine and synthetic frames, yet the u factor would only properly occur in pristine data while the frame interpolation methods each would generate an over-smooth average u. Thus it is reasonable for the difference between the natural u and the average u to be greater than the difference between two average us generated by the two synthesis methods. To use this behavior for interpolation detection, for each frame, the interpolated parallel can be generated from the frame before and after with any other good interpolation method that fits the aforementioned description. Next, the prediction error can be measured as the difference between the interpolated frame and the observed one. Consequently, this difference can be used for distinguishing pristine frames from interpolated ones by using a distance measure. Alternatively, this prediction error *image* can be fed to a classifier which specializes in the detection of interpolated frames for better performance.

3 Experiment Setup

We provide the large-scale Morph Cut dataset for the task of unit-selection based facial video manipulation detection training and testing on which we empirically verify the detection hypothesis. Furthermore, in our benchmark we perform the detection task with four applicable detection methods from the literature. The details of the dataset along with the experiment setup is explained in the following.

3.1 Morph Cut Dataset Details

The VoxCeleb2 [NCZ17] dataset is used as a basis for video selection, which contains a collection of interview videos from celebrities hosted on the video-sharing platform YouTube. The videos are ranked based on the face bounding-box movements, and on the suitable videos, uniform random sampling is applied to select candidate points for morph-cut. Next, the candidates with high structural similarity index [WB09] are selected and two morph-cut transitions are automatically added to each video using Extendscript. The Morph Cut dataset contains $1,000$ videos with an average duration of 2.75 seconds. This dataset adds up to $\sim 83,000$ frames with $\sim 27,500$ morphed frames and a ratio of 33% morphed frames to pristine ones. The videos are split three sets corresponding to training, validation, and the test data according to numbers in Table 1. The video parameters are summarized in Table 2. The videos are accompanied by frame-level labels corresponding to whether each frame is morphed or pristine. All reported results are based on frame-level classification performance between the morphed frames and the pristine ones.

Set	Count
Train	700
Dev	150
Test	150

Tab. 1: The number of videos in each set of the constructed Morph Cut dataset.

Video parameters
MPEG-4 (Base Media / Version 2)
480p (854×480)
30 FPS (Frames-Per-Second)
AVC (NTSC)

Tab. 2: The parameters used to create each video in the constructed Morph Cut dataset.

3.2 Proposed Detector

For the detector's reference frame-interpolation method, the pre-trained CyclicGen [Li19] convolutional neural network is used. For a given pair of frames, this network produces a high-quality intermediate interpolated frame. Using this network, for each frame in a video, a corresponding interpolated frame is synthesized based on the frame before and after, and the prediction error is calculated in terms of a difference image. The resulting prediction error *images* on cropped face regions are then converted to gray-scale and fed to a simple convolutional neural network for frame-level classification. The input to the

network is augmented with the *context* prediction error images of two frames before and after, resulting in an input shape of $64 \times 64 \times 5$. The training and evaluation pipeline is visualized in Figure 2 and the classifier network architecture is summarized in Table 3.

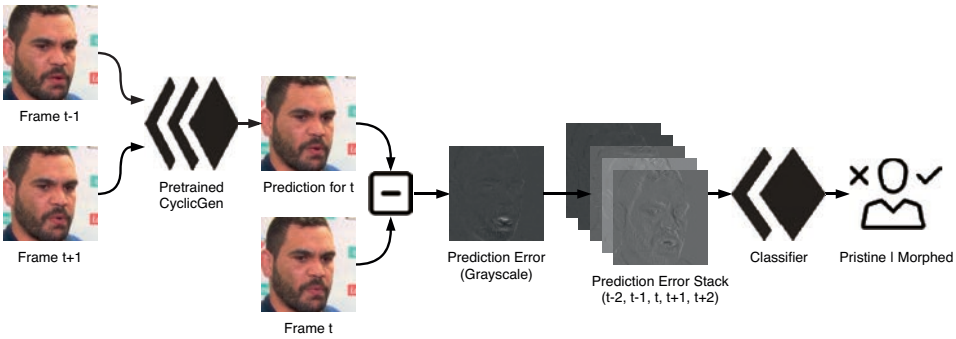


Fig. 2: The training and evaluation pipeline in the proposed method.

Layer	Output Shape	Parameters
Conv2D	(62, 62, 128)	Kernel=(3,3)
MaxPooling2D	(31, 31, 128)	Pool=(2,2)
Conv2D	(29, 29, 128)	Kernel=(3,3)
MaxPooling2D	(14, 14, 128)	Pool=(2,2)
Conv2D	(12, 12, 256)	Kernel=(3,3)
MaxPooling2D	(6, 6, 256)	Pool=(2,2)
Conv2D	(4, 4, 512)	Kernel=(3,3)
MaxPooling2D	(2, 2, 512)	Pool=(2,2)
Flatten	(2048)	
Dense	(512)	
DropOut	(512)	
Dense	(2)	

Tab. 3: The network architecture of the classifier. The network contains $1.6M$ trainable parameters.

3.3 Baseline Methods

For baseline methods to be used in our benchmark, we relied on recently published and reproducible detection methods for face-morph detection [AJ17], time-aware Deepfake detection [GD18], inter-frame forgery detection [BN18], and general purpose image classification [Ch17]. Among the four methods, [GD18] and [BN18] utilize temporal information while [AJ17] and [Ch17] rely only on static face images. All methods provide frame-level decision.

The first method is based on topological data analysis for image tampering detection described in the paper of the same name [AJ17]. This method was originally created to detect morphing attacks on face images by extracting features from the texture of the image itself, making the method sensitive to image tampering through the degradation of the image. For this method, we first extract the cropped faces from each frame in the dataset and construct

a 1-skeleton of the full rips simplicial complex for each face image, which is then fed into an SVM classifier to attempt and recognize the morphed faces against the pristine ones.

The second method relies on recurrent neural networks for Deepfake detection [GD18]. The cropped face images are used as input to the network and all parameters are kept the same as described in the paper except we are training with fewer epochs. The third method relies on 3D convolutional neural networks for the detection of inter-frame forgery as described in [BN18]. Finally, due to the outstanding performance of the Xception-Net [Ch17] for Deepfake detection task, the pre-trained network is fine-tuned on the task of morph-cut detection on individual images.

4 Results and Discussion

Table 4 summarizes the detection accuracy of the proposed method in comparison to the baseline methods. The proposed method achieves the highest detection accuracy of 95.1% on the test set, followed surprisingly by the fine-tuned XceptionNet at 77.0%. The other three baseline methods show limited success in the detection of morph-cut frames. The detection-error-tradeoff (DET) curve for the top 3 best-performing methods is shown in Figure 3. In this figure, APCER stands for attack presentation classification error rate and BPCER stand for bona fide presentation classification error rate, which correspond to the missed detection and the false alarm rate of a biometric presentation attack detection system respectively following the ISO/IEC 30107 standard terminology[11]. The proposed method achieves an acceptable detection equal-error-rate (EER) of 4.95%.

Method	**Test Accuracy**
Topological Data Analysis [AJ17]	50.2%
Deepfake Video Detection [GD18]	59.0%
Inter-Frame Forgery C3D [BN18]	67.4%
Fine-tuned XceptionNet [Ch17]	77.0%
Proposed Method	95.1%

Tab. 4: The detection accuracy of the proposed method in comparison to the baseline methods. The results show the frame-level performance.

Examples of the prediction errors which are used as input to the classifier in the proposed method are visualized in Figure 4. Natural variations are clearly visible in prediction errors in pristine frames, while these variations are not observed in the morphed (interpolated) ones. Figure 5 shows the probability density distribution of average prediction error per frame over pristine and morphed frames. The morphed frame average prediction error distribution is shifted towards zero compared to the pristine distribution, confirming the hypothesis proposed in Section 2.2. The clear distinction between the pristine and morphed frame prediction errors visualized in Figure 4 and 5 show the effectiveness of prediction error *images* in isolating useful features for morphed face detection.

[11] `https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-3:ed-1:v1:en`

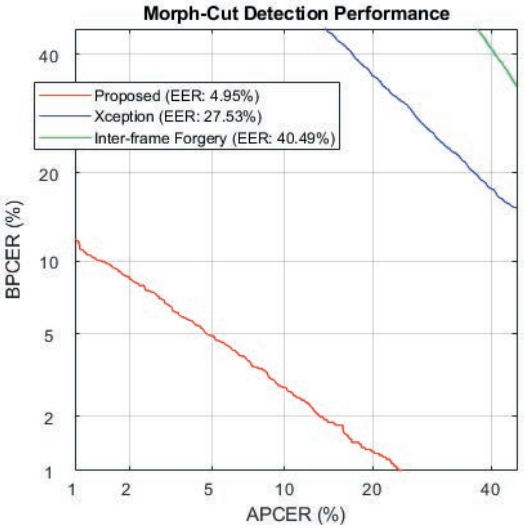


Fig. 3: The DET curve for the frame-level detection performance of the proposed method, the fine-tuned Xception-Net[Ch17], and the inter-frame forgery detection method[BN18]. The equal-error-rate (EER) value for the aforementioned methods is shown in the figure legend.

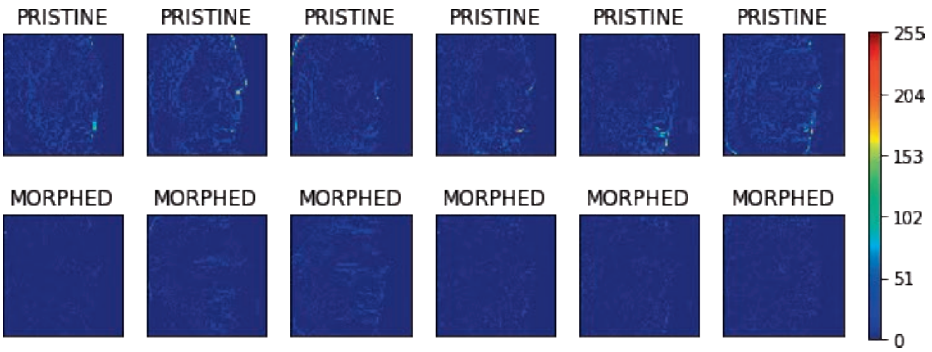


Fig. 4: Example of prediction error *images* of cropped faces in a six-frame sequence of pristine frames (top) and morph-cut frames (bottom) in a video. The images visualize the absolute gray value difference per pixel between the interpolation output and the actual frame.

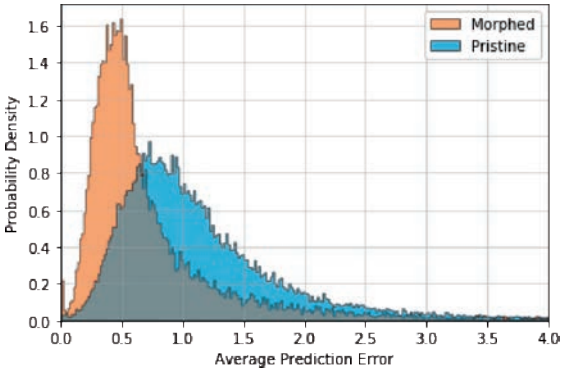


Fig. 5: The probability density distribution of average prediction error per frame for pristine and morphed frames across the dataset.

5 Conclusion

In this article, we addressed the problem of unit-selection based facial video manipulation by providing the first large-scale dataset of videos manipulated by popular video-editing software. Furthermore, we proposed a detection method that relies on frame-interpolation prediction-errors as discriminative features for the detection of morphed frames. The proposed method outperforms the baseline methods by a wide margin. The high frame-level performance of the proposed method shows its capacity in reliably detecting unit-selection based video manipulation and confirms the detection hypothesis that synthetic frames demonstrate higher similarity to each other than to pristine ones.

References

[AJ17] Asaad, Aras; Jassim, Sabah: Topological data analysis for image tampering detection. In: International Workshop on Digital Watermarking. Springer, pp. 136–146, 2017.

[Am19] Amerini, Irene; Galteri, Leonardo; Caldelli, Roberto; Del Bimbo, Alberto: Deepfake Video Detection through Optical Flow Based CNN. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. Oct 2019.

[ASAS16] Al-Sanjary, Omar Ismael; Ahmed, Ahmed Abdullah; Sulong, Ghazali: Development of a video tampering dataset for forensic investigation. Forensic Science International, 266:565 – 572, 2016.

[BCS97] Bregler, Christoph; Covell, Michele; Slaney, Malcolm: Video Rewrite: Driving Visual Speech with Audio. In: SIGGRAPH. SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., USA, p. 353–360, 1997.

[BLA12] Berthouzoz, Floraine; Li, Wilmot; Agrawala, Maneesh: Tools for Placing Cuts and Transitions in Interview Video. ACM Trans. Graph., 31(4), July 2012.

[BN18] Bakas, Jamimamul; Naskar, Ruchira: A Digital Forensic Technique for Inter–Frame Video Forgery Detection Based on 3D CNN. In: International Conference on Information Systems Security. Springer, pp. 304–317, 2018.

[Ch17] Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807, 2017.

[CJS12] Chao, Juan; Jiang, Xinghao; Sun, Tanfeng: A Novel Video Inter-Frame Forgery Model Detection Scheme Based on Optical Flow Consistency. In: IWDW. IWDW'12, Springer-Verlag, Berlin, Heidelberg, p. 267–281, 2012.

[Da11] Dale, Kevin; Sunkavalli, Kalyan; Johnson, Micah K.; Vlasic, Daniel; Matusik, Wojciech; Pfister, Hanspeter: Video Face Replacement. In: SIGGRAPH Asia. SA '11, Association for Computing Machinery, New York, NY, USA, 2011.

[GD18] Güera, D.; Delp, E. J.: Deepfake Video Detection Using Recurrent Neural Networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6, 2018.

[JE19] Johnston, Pamela; Elyan, Eyad: A review of digital video tampering: From simple editing to full synthesis. Digital Investigation, 29:67 – 81, 2019.

[KRB19] Khodabakhsh, A.; Ramachandra, R.; Busch, C.: Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6, 2019.

[Li19] Liu, Yu-Lun; Liao, Yi-Tung; Lin, Yen-Yu; Chuang, Yung-Yu: Deep Video Frame Interpolation using Cyclic Frame Generation. In: Proceedings of the 33rd Conference on Artificial Intelligence (AAAI). 2019.

[MV15] Mattheyses, Wesley; Verhelst, Werner: Audiovisual speech synthesis: An overview of the state-of-the-art. Speech Communication, 66:182 – 217, 2015.

[NCZ17] Nagraniy, Arsha; Chungy, Joon Son; Zisserman, Andrew: VoxCeleb: A large-scale speaker identification dataset. INTERSPEECH, 2017-August:2616–2620, 2017.

[Sa19] Sabir, Ekraam; Cheng, Jiaxin; Jaiswal, Ayush; AbdAlmageed, Wael; Masi, Iacopo; Natarajan, Prem: Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In: CVPR Workshops. June 2019.

[Sc19] Scherhag, U.; Rathgeb, C.; Merkle, J.; Breithaupt, R.; Busch, C.: Face Recognition Systems Under Morphing Attacks: A Survey. IEEE Access, 7:23012–23026, 2019.

[SM16] Sitara, K.; Mehtre, B.M.: Digital video tampering detection: An overview of passive techniques. Digital Investigation, 18:8 – 22, 2016.

[Ve20] Verdoliva, Luisa: Media forensics and deepfakes: an overview. arXiv preprint arXiv:2001.06564, 2020.

[WB09] Wang, Zhou; Bovik, Alan C.: Mean squared error: Lot it or leave it? A new look at signal fidelity measures. IEEE Signal Processing Magazine, 26(1):98–117, 2009.

[WKG18] Wandzik, L.; Kaeding, G.; Garcia, R. V.: Morphing Detection Using a General- Purpose Face Recognition System. In: 2018 26th European Signal Processing Conference (EUSIPCO). pp. 1012–1016, 2018.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fuŕ Informatik, Bonn 2020 97

Longitudinal study of voice recognition in children

Sandip Purnapatra[1], Priyanka Das[2], Laura Holsopple[3], Stephanie Schuckers[4]

Abstract: Speaker recognition as a biometric modality is on the rise in the consumer marketplace for banking, online services, and personal assistant services with a potential for wider application areas. Most current applications involve adults. One of the biggest challenges in speaker recognition for children is the change in the voice properties as a child age. This work proposes a baseline longitudinal dataset from the same 30 children in the age group of 4 to 14 years over a time frame of 2.5 years and evaluates speaker recognition performance in children with the available speaker recognition technology.

Keywords: Speaker verification, Children's voice, MFCC, LFCC, GMM, JFA, ISV, Inter-session variability.

1 Introduction

Biometric recognition has proliferated in the last two decades with applications in government (border security, immigration, identity at birth, distribution of benefits, refugee efforts) and consumer market (e-commerce, banking, healthcare). Biometric recognition based on voice uses unique features in the speaker's voice to ascertain identity [Ma00]. Voice biometrics uses acoustic properties specific to individual subjects and can be used in situations involving virtual presence over any telephone or internet. Voice biometrics is applied mostly for speaker verification. Speaker verification can be text-dependent or text-independent. For either, the biometric characteristic contains features of the voice specific to a person. Voice biometrics for speaker recognition has been used sparsely since late 1990s. However, in the past decade the application of speaker recognition proliferated in the consumer market for personal assistant services in mobile devices, online services requiring authentication like online banking services, call centers and other services.

Most of the prior research involving voices of children are based on physiological changes of voices with targeted applications like gender recognition, and speech recognition. Speaker recognition performance is still a relatively unexplored research area. One of the few studies in this area shows that as a child ages, their vocal properties changes, impacting the performance of speaker recognition [SRJ18]. The paper is described in more detail at the end of this section.

Studies have shown that developmental speech production, especially vocal tract growth, introduces age-dependent spectral and temporal variability in the speech signal of children. Such variability evoke challenges for robust automatic recognition of children's speech [PN03]. However, no research regarding the influence of vocal tract growth for automatic speaker recognition has been

[1] PhD Student, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, purnaps@clarkson.edu
[2] PhD Student, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, prdas@clarkson.edu
[3] Associate director of CITeR, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, lholsopp@clarkson.edu
[4] Professor, Electrical and Computer Engineering, Clarkson University, 8 Clarkson Avenue, Potsdam, NY-13676, US, sschucke@clarkson.edu

performed for children's voices. According to [Ma00], changes in the voice properties in children add to the basic challenges in voice recognition- background and channel noise; variable and inferior microphones and telephones; and extreme hoarseness, fatigue, or vocal stress.

Extraction of useful information from speech has been researched actively in the last three decades. Mel frequency cepstral coefficient (MFCC), which mimics the frequency response of the human ear, is a well established feature used extensively in most voice/speaker recognition techniques [MBE10]. MFCC filters are designed in accordance to the critical bandwidth frequencies that the human ear perceives. MFCC uses two types of filter- linearly spaced and logarithmically spaced [NS14]. Linear frequency cepstral coefficients (LFCC) is another feature extraction technique that uses only linearly spaced filters. LFCC provides equal details for all frequencies [Re94]. In the higher frequency region of speech, LFCC uses higher number of filterbank compared to MFCC. Inter to intra class speaker variability ratio or f-ratio is significantly higher in LFCC than MFCC [LL09].

In the 1990's many methods like simple template matching, statistical pattern recognition, dynamic time-warping methods were used for speaker recognition. Hidden markov models (HMM), Gaussian mixture model (GMM), universal back ground model (UBM) and multi-layer perceptron gained popularity as speaker recognition techniques in the early 2000's [NS14]. In the last decade, speaker recognition techniques that are based on different types of factor analysis i.e. joint factor analysis (JFA), i-vectors, linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (PLDA) produced improved speaker recognition results [Ka14]. GMM based speaker recognition techniques are not designed to compensate for the inter-session sound variability of different recordings and fails to minimize the the variation in enrollment and probe recordings induced by environmental factors. JFA minimizes these season variability caused by the sound difference of a given speaker's different recordings [Ke07a] [Mc10]. JFA is the Gaussian distribution of HMM supervectors which are speaker and channel dependent and account a few hidden variables of speaker and channel factors or high dimensional GMM super-vectors. JFA model assumes the speaker factor in two different recordings remain same but the channel factor or the recording environment varies from session to session [Ke05] [Ke07a]. JFA models does not work on the speaker verification on short utterance recordings ($<$10 seconds). Rather than modelling the speaker or channel variability space, intermediate-size vector or i-vector models speaker and channel variability in a low dimensional, single total-variability space that can map the utterances (short utterances as well) of the speakers and help convert the speaker recognition problem from a high dimensional to a low dimensional one [Ka14] [De09]. Inter-session variability (ISV) is another modelling approach similar to JFA which aims to reduce inter-session variability in GMM speaker model space [VS08] [Ke07b].The main difference between ISV and JFA is, while ISV modelling assumes that *within subject* variability is dominant in the linear domain of the GMM super vector low-dimensional subspace, JFA assumes that the *between-subject* variability is contained in the low-dimensional subspace [Wa11]. i-vector is also designed to mitigate the speaker variability caused by collecting data from different sources. As i-vectors are computed from the hidden variables of the factor analysis model, it requires huge amount of training data. However, i-vector does not address channel variability; it needs to be combined with other models such as LDA, probabilistic LDA (PLDA), cosine similarity scoring (CSS), within-class covariance normalization (WCCN), which divides the total variability space into session and speaker variability sub-spaces [De09] [Ka11], to mitigate the discrepancy between channel noise of different samples. Deep learning (DL) based speaker recognition systems have the capacity to extract the low-dimensional features and achieve strong speaker recognition performance [Gu20] [Li20b]. Although the DL models produce an improved speaker recognition performance compared to classifiers that require hand-crafted features, they are more complex, requires massive amount of labeled training data, has high computation and storage cost [Li20a].

There has been limited work on speaker verification in children. Safavi et al. [Sa16] [Sa14] [SRJ18] performed automatic speaker, gender and age group identification of approximately 1100 children of different age groups using MFCC, delta and delta-square features and GMM-UBM, GMM-SVM, i-vector-PLDA based models to achieve maximum 99% identification accuracy using only 10 seconds speech recording. However, the dataset details for speaker recognition analysis did not mention

a multi-session collection. More study is needed where the enrollment and verification happen on different days and in multiple sessions spaced by considerable time gap, particularly as a child ages. To the best of our knowledge no report has been published on longitudinal voice biometric recordings for speaker recognition in children. Our study is the first work evaluating speaker verification performance in children with data collected over multi-sessions. In this study, we analyzed longitudinal speaker verification performance in children over a period of 2.5 years with data collected from six sessions with inter-session gap of six months, in time frames of 6, 12, 18, 24 and 30 months between enrollment and verification samples, for the age group of 4 to 14 years using the available technology for adult speaker verification, with approximate recording duration of 90 seconds per subject per session. We report on the longitudinal robustness of speaker verification in children as they age. This work contributes to the research domain by-

1. providing a baseline longitudinal dataset for speaker recognition in children to advance research in this field;
2. evaluating the robustness of established techniques for speaker recognition with child voice data;
3. analyzing the longitudinal speaker verification performance in children.

The rest of the paper is organized in four sections- Section 2 explains data collection protocol, Section 3 details the experimentation steps, Section 4 highlights the results achieved and Section 5 provides a discussion of the limitations of this study, future scopes and concludes on the feasibility of the state-of-art techniques of voice recognition in children.

2 Dataset

The dataset consists of data from the same 30 subjects, for over 2.5 years period, collected from six sessions at an approximated interval of six months from subjects aged between 4 and 11 yrs at enrollment. Using the first session data as the enrollment, longitudinal performance of the dataset has been tested for five subsequent time instances at 6, 12, 18,24 and 30 months. Subject count for each enrollment age between 4 and 11 years are 1, 1, 5, 3, 6, 2, 8 and 4, respectively. The data used for this study is part of a multi-modal biometric dataset collected from the same children for research purpose in cooperation with a local school. The research team sets up collection stations at the school every six months for the collection days using the same equipment. The collection room may vary based on availability which may impact the data.

Voice data is collected at a sampling rate of 44.1 KHz using a microphone by Audio-Technica with frequency response 20Hz to 16KHz and bit depth of 16bit and a publicly available software, Audacity. At each session the subjects are prompted by a series of images to speak simple common words like numbering (1-10), name of animals and common objects known to children and at the end they are asked to describe a scene displayed to them as an image. The speech duration varies based on the speaking speed of the subjects including pauses in between words. Only one sample is collected at each session from each subject of approximate duration of 90 seconds. The protocol and the content was same in all collections. However, the order of images, and thus the words, may have varied. This study focuses on text-independent verification and the content and the order are not considered. Since the data is collected in a school environment, even with our best effort, the collected data have inconsistent noise ranging from sound of people walking, opening or closing of doors, and people talking nearby.

3 Experimentation

3.1 Experimentation Platform: Bob

Bob [An12] [An17] is an open source, reproducible signal processing toolbox. *bob.bio.spear* is a speaker recognition package in the Bob platform having supporting tools for speech data pre-processing, feature extraction, matching and analysis. All experiments with our child voice data has been performed in this platform.

3.2 Data Pre-processing

The data collected for this experiment is in a real life scenario i.e, the data includes channel noise from devices and other environmental noise. Practical applications may not include noise free environment. Thus, it is important to pre-process the data without losing the voice print and distorting the features Our pre-processing includes three steps:

1. Band pass filtering between range 125 Hz and 8000 Hz
2. Mean-Variance Normalization
3. Silence Removal by identifying- (a) Short-time Energy and (b) Spectral Centroid

These data pre-processing steps were performed in MATLAB 2019a, prior to our experimentation in Bob. The Bob experimentation also includes data pre-processing. For our experimentation we used the inbuilt pre-processing resource- *energy_thr*[ID], which is a thresholded energy based voice detection function. The default threshold is 15% of the maximum energy of the input signal, which was used for the secondary pre-processing of our data in the Bob platform. No data was removed due to quality or noise purpose before experimentation in the Bob platform. However, the pre-processor used in the bob-platform failed to process eight samples from eight different subjects at random sessions.

3.3 Feature Extraction and Algorithm

State of art features and algorithms were tested to assess longitudinal speaker recognition performance in children over 2.5 years. Two different feature sets- MFCC and LFCC, were tested with 20 and 60 coefficients for both the feature extraction techniques. Three algorithms- GMM, ISV and JFA, were used to assess performance. Speaker recognition performance from 12 different feature-algorithm combinations tested for our study are tabulated in Table 1.

4 Results and Analysis

Performance is evaluated in terms of False Accept Rate (FAR), False Reject Rate (FRR) and Equal Error Rate (EER). Figure 1 - 12 shows the score distributions and Figure 13 - 24 shows the ROCs for 12 different feature-algorithm combinations for each five longitudinal time instances (6,12,18,24 and 30 months) for the same 30 subjects. Table 1 summarizes the performance at each time instances for each of 12 combination of feature-algorithm in terms of EER.

With MFCC60 and LFCC60, there is decaying variability in the score distribution with ISV algorithm (refer Fig. 4, 10). ISV was reported in literature to have improved speaker verification performance in adults [VS08]. However, we note a drastic degradation in performance with our children

Tab. 1: Speaker Verification performance

Feature	Algorithm	EER (%) 6 month	EER (%) 12 month	EER (%) 18 month	EER (%) 24 month	EER (%) 30 month
MFCC 20	GMM	22	26	30	24	42
MFCC 20	ISV	48	46	56	52	54
MFCC 20	JFA	34	38	35	40	43
MFCC 60	GMM	36	38	40	43	42.5
MFCC 60	ISV	36	44	40	46	46
MFCC 60	JFA	43	37	44	46	52
LFCC 20	GMM	26	34	29	40	48
LFCC 20	ISV	48	47	50	59	56
LFCC 20	JFA	43	38	45	44	50
LFCC 60	GMM	38	35	41	45	51
LFCC 60	ISV	48	52	46	52	54
LFCC 60	JFA	44.5	36	42	52	47.5

dataset as reflected in the ROCs (refer Fig. 15, 16, 21, 22). Though the performance improves for 60 feature dimension compared to 20 feature dimension, the performance of ISV is poor compared to both JFA and GMM (refer 1).

Joint Factor Analysis, which is an extension of ISV, is designed to reduce inter-session variability for intra-subject data and to reduce the high enrollment requirement. The reduced inter-session variability is reflected in the score distributions in Figure 5, 6, 11 as well as in the reduced variability in the performance between longitudinal time instances (6, 12, 18, 24 and 30 months). However, the overall performance is poor compared to GMM with the same set of features.

MFCC20, MFCC60, LFCC20 and LFCC60 features has high variability across longitudinal time instances (6,12,18,24 and 30) with GMM. There is a distinct decay in genuine match scores with GMM for MFCC20 and LFCC20 (refer Fig. 1, 7). The score distribution for 60 dimensional features for MFCC and LFCC show higher variability. However, GMM performs best with all 4 configurations-MFCC-20, MFCC60, LFCC-20, LFCC-60, compared to ISV and JFA. The best performance is observed for the MFCC20 and GMM combination in terms of FAR and FRR (EER varies from 22% at 6 month time instance to 42% at 30 month time instance) compared to other algorithms and features. Overall, 20 dimensional feature vector for both MFCC and LFCC perform better compared to 60 dimensional features. Almost all feature-algorithm combination fails to perform at 30 month time frame with EER ranging between 42% to 56 %.

5 Discussion, Limitation and Future Scope

In the last few years several speaker verification systems has been proposed. However, impact of increased time between voice enrollment and probe samples on speaker recognition performance are still an unexplored area, especially in children. This work is an attempt to answer the question- *Are the available voice recognition techniques robust enough to recognize children as they age?* For this purpose, a dataset has been collected from the same 30 children in six sessions over 2.5 years. The data has been analyzed using state of art features and algorithms that have proved effective for adult speaker verification.

From our analysis, we conclude that MFCC20 features and GMM algorithm performs best for longitudinal speaker verification in children. However, the best performance is not on par with the

Score distribution for genuine matches as time increases between enrollment and verification

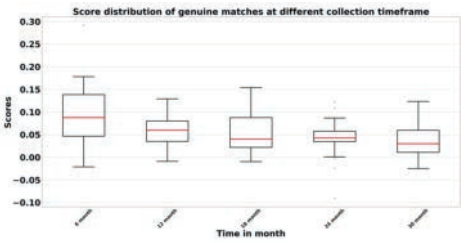


Fig. 1: Feature: MFCC20; Algo: GMM

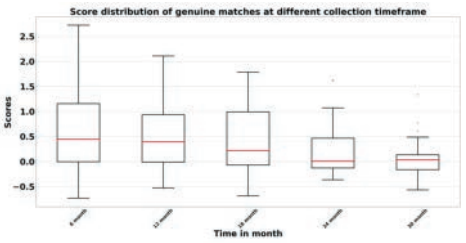


Fig. 2: Feature: MFCC60; Algo: GMM

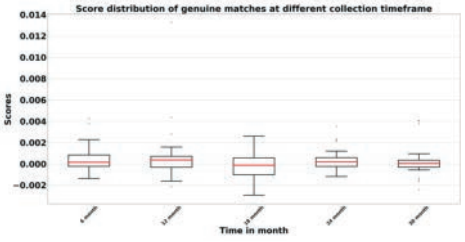


Fig. 3: Feature: MFCC20; Algo: ISV

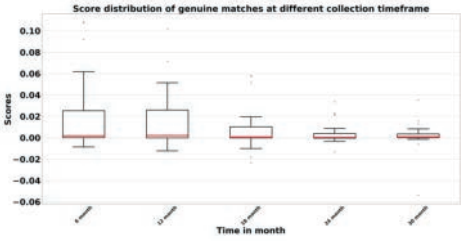


Fig. 4: Feature: MFCC60; Algo: ISV

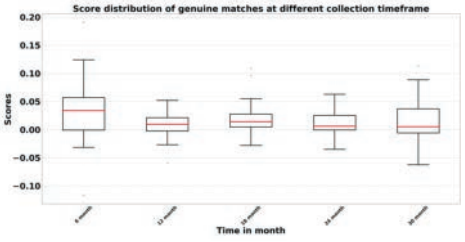


Fig. 5: Feature: MFCC20; Algo: JFA

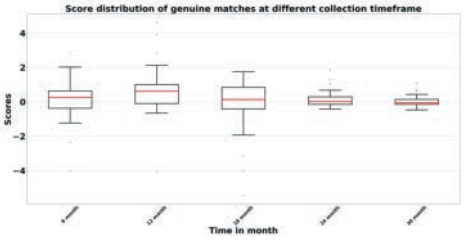


Fig. 6: Feature: MFCC60; Algo: JFA

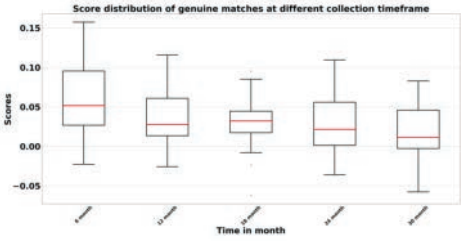


Fig. 7: Feature: LFCC20; Algo: GMM

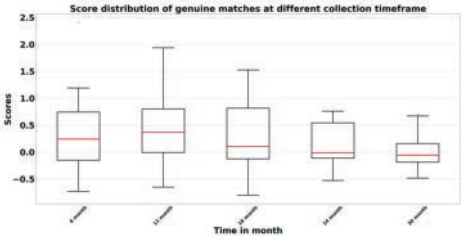


Fig. 8: Feature: LFCC60; Algo: GMM

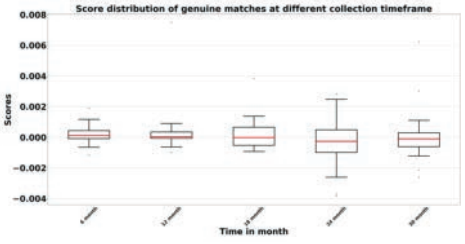


Fig. 9: Feature: LFCC20; Algo: ISV

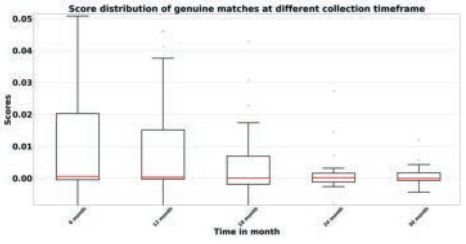


Fig. 10: Feature: LFCC60; Algo: ISV

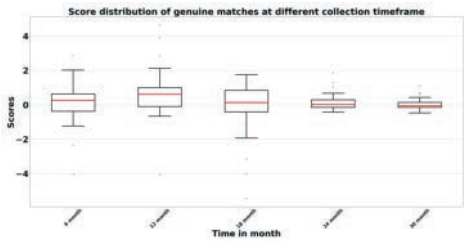


Fig. 11: Feature: LFCC20; Algo: JFA

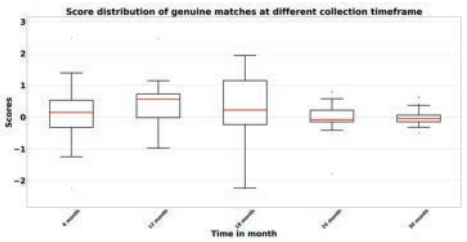


Fig. 12: Feature: LFCC60; Algo: JFA

expected biometric recognition performance. The state of art algorithms (ISV and JFA) designed to reduce inter session variability and improve recognition performance, do perform well in children. However, these are not commercially developed algorithms, which we assume might perform differently. We note that there is need for improvement of speaker recognition in children with the development of appropriate features and algorithms.

The data used in the study was collected in a real life scenario with background noise including sound of people walking by, talking, opening and closing of doors and other miscellaneous noise. However, not all data for all subjects have noise and the noise level varies between sessions and subjects. No complete session was deleted due to noise. We removed pauses between utterances to reduce noise in the data. Most noise frequencies are in the range of human voice frequencies. Thus, even with the best effort it was not possible to eliminate noise frequencies from the signal without effecting the voice properties. Thus it is expected to have degraded performance in recognition compared to ideal voice samples. To the best of our knowledge, no publicly available multi-session voice dataset from children is available to support research in this field. Pre-trained networks on adult data has proved inefficient when used in applications involving children for other modalities like face, where high variability is observed with aging [DNJ18]. However, it can be a work for future to test the viability of such approach with child speaker verification. State of art algorithms with hand crafted features do not require a large amount of data for supervised training has proved high efficiency. Cases with limited amount of data needs robust algorithm pipeline for applications in terms of both features and algorithm. We recognize that non-availability of dataset is a hindrance to our research community. We also recognize privacy and sensitivity related to child biometric data. We are in the process of sharing our dataset through BEAT platform to support research in this field while protecting data privacy. All algorithms used for analysis are also available through an interface from BEAT to the Bob platform.

This work initiates research in the field of child voice recognition impacted by aging. For future work statistical modelling of the variation in voice signature features may help in modelling biometric aging in child voices. The very basis of biometrics is temporal-stability. Time-invariant voice features need to be defined for child in order to be useful for biometric applications. Research on robust feature and classification techniques are required to address speaker recognition with intra-class variability due to aging in children. Further research in this field is needed to support widespread application of voice biometrics across all age groups. We conclude that the state of art algorithms for speaker recognition performance in adults does not reflect similarly in the case of speaker recognition in children for the age group of 4 to 14 years. There is a need for development of age-independent features and algorithms for child speaker recognition for longitudinal biometric applications.

6 Acknowledgement

This research was funded by Center for Identification Technology and Research (CITeR) and National Science Foundation (NSF)(Grant # :1650503). The database creation was made possible by

ROCs as time increases between enrollment and verification with different feature sets and algorithms

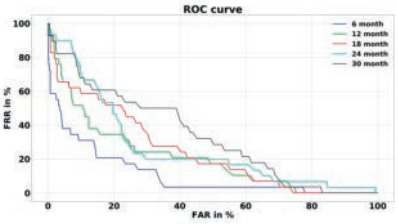


Fig. 13: Feature: MFCC20; Algo: GMM

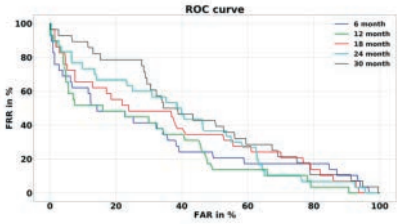


Fig. 14: Feature: MFCC60; Algo: GMM

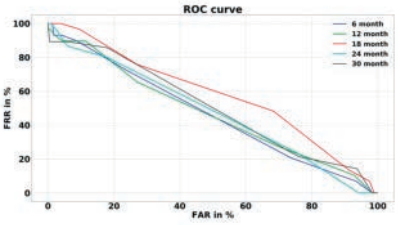


Fig. 15: Feature: MFCC20; Algo: ISV

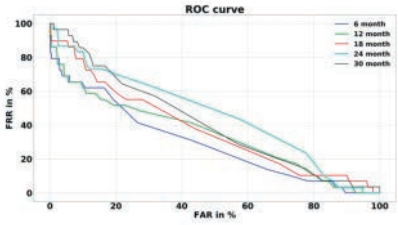


Fig. 16: Feature: MFCC60; Algo: ISV

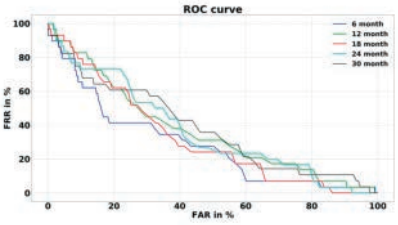


Fig. 17: Feature: MFCC20; Algo: JFA

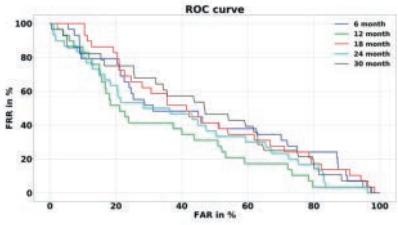


Fig. 18: Feature: MFCC60; Algo: JFA

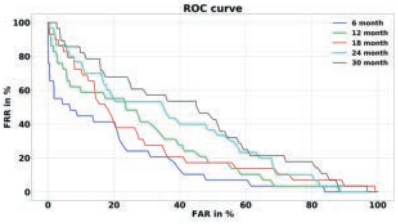


Fig. 19: Feature: LFCC20; Algo: GMM

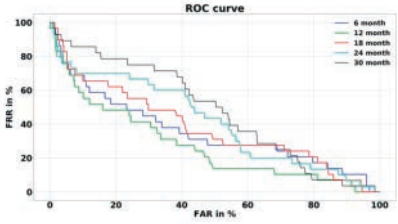


Fig. 20: Feature: LFCC60; Algo: GMM

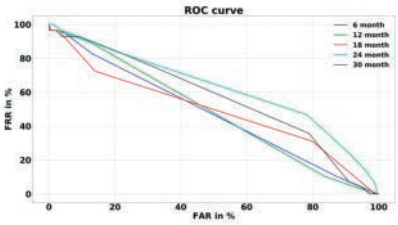


Fig. 21: Feature: LFCC20; Algo: ISV

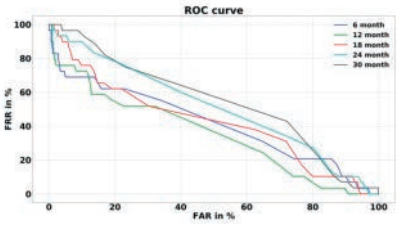


Fig. 22: Feature: LFCC60; Algo: ISV

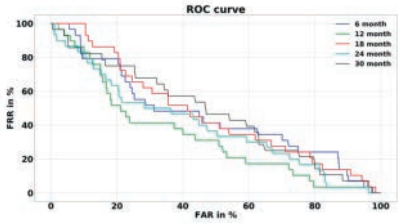


Fig. 23: Feature: LFCC20; Algo: JFA

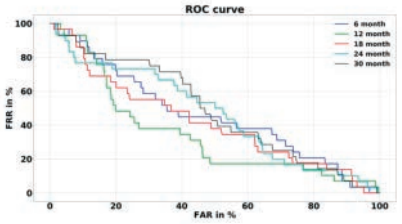


Fig. 24: Feature: LFCC60; Algo: JFA

voluntary participation of all enrolled subjects in the study, their parents/guardians and the hard work of the data collecting team from Clarkson University. We would also extend our gratitude to the Potsdam Elementary School and Potsdam Middle School administration and staff members for their continued support in academic research.

References

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

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

[De09] Dehak, Najim; Dehak, Reda; Kenny, Patrick; Brümmer, Niko; Ouellet, Pierre; Dumouchel, Pierre: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Tenth Annual conference of the international speech communication association. 2009.

[DNJ18] Deb, Debayan; Nain, Neeta; Jain, Anil K: Longitudinal study of child face recognition. In: 2018 International Conference on Biometrics (ICB). IEEE, pp. 225–232, 2018.

[Gu20] Gusev, Aleksei; Volokhov, Vladimir; Andzhukaev, Tseren; Novoselov, Sergey; Lavrentyeva, Galina; Volkova, Marina; Gazizullina, Alice; Shulipa, Andrey; Gorlanov, Artem; Avdeeva, Anastasia et al.: Deep speaker embeddings for far-field speaker recognition on short utterances. arXiv preprint arXiv:2002.06033, 2020.

[ID] IDIAP: , Energy Theshold Pre-processor. https://pydoc.net/bob.bio.spear/3.1.0/bob.bio.spear.preprocessor.Energy_Thr/. Accessed: 2020-08-17.

[Ka11] Kanagasundaram, Ahilan; Vogt, Robbie; Dean, David B; Sridharan, Sridha; Mason, Michael W: I-vector based speaker recognition on short utterances. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association. International Speech Communication Association (ISCA), pp. 2341–2344, 2011.

[Ka14] Kanagasundaram, Ahilan; Dean, David; Sridharan, Sridha; Gonzalez-Dominguez, Javier; Gonzalez-Rodriguez, Joaquín; Ramos, Daniel: Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. Speech Communication, 59:69–82, 2014.

[Ke05] Kenny, Patrick: Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montreal,(Report) CRIM-06/08-13, 14:28–29, 2005.

[Ke07a] Kenny, Patrick; Boulianne, Gilles; Ouellet, Pierre; Dumouchel, Pierre: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, 15(4):1435–1447, 2007.

[Ke07b] Kenny, Patrick; Boulianne, Gilles; Ouellet, Pierre; Dumouchel, Pierre: Speaker and session variability in GMM-based speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 15(4):1448–1460, 2007.

[Li20a] Li, Ruirui; Jiang, Jyun-Yu; Li, Jiahao Liu; Hsieh, Chu-Cheng; Wang, Wei: Automatic speaker recognition with limited data. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 340–348, 2020.

[Li20b] Liang, Tianyu; Liu, Yi; Xu, Can; Zhang, Xianwei; He, Liang: Combined Vector Based on Factorized Time-delay Neural Network for Text-Independent Speaker Recognition. In: Proc. Odyssey 2020 The Speaker and Language Recognition Workshop. pp. 428–432, 2020.

[LL09] Lei, Howard; Lopez, Eduardo: Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In: Tenth Annual Conference of the International Speech Communication Association. 2009.

[Ma00] Markowitz, Judith A: Voice biometrics. Communications of the ACM, 43(9):66–73, 2000.

[MBE10] Muda, Lindasalwa; Begam, Mumtaj; Elamvazuthi, Irraivan: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083, 2010.

[Mc10] McLaren, Mitchell; Vogt, Robert; Baker, Brendan; Sridharan, Sridha: A comparison of session variability compensation approaches for speaker verification. IEEE Transactions on Information Forensics and Security, 5(4):802–809, 2010.

[NS14] Nijhawan, Geeta; Soni, MK: Speaker recognition using support vector machine. International Journal of Computer Applications, 87(2), 2014.

[PN03] Potamianos, Alexandros; Narayanan, Shrikanth: Robust recognition of children's speech. IEEE Transactions on speech and audio processing, 11(6):603–616, 2003.

[Re94] Reynolds, Douglas A: Experimental evaluation of features for robust speaker identification. IEEE Transactions on Speech and Audio Processing, 2(4):639–643, 1994.

[Sa14] Safavi, Saeid; Najafian, Maryam; Hanani, Abualsoud; Russell, Martin J; Jancovic, Peter: Comparison of speaker verification performance for adult and child speech. In: WOCCI. pp. 27–31, 2014.

[Sa16] Safavi, Saeid; Najafian, Maryam; Hanani, Abualsoud; Russell, Martin J; Jancovic, Peter; Carey, Michael J: Speaker recognition for children's speech. arXiv preprint arXiv:1609.07498, 2016.

[SRJ18] Safavi, Saeid; Russell, Martin; Jančovič, Peter: Automatic speaker, age-group and gender identification from children's speech. Computer Speech & Language, 50:141–156, 2018.

[VS08] Vogt, Robbie; Sridharan, Sridha: Explicit modelling of session variability for speaker verification. Computer Speech & Language, 22(1):17–38, 2008.

[Wa11] Wallace, Roy; McLaren, Mitchell; McCool, Christopher; Marcel, Sebastien: Inter-session variability modelling and joint factor analysis for face authentication. In: 2011 International Joint Conference on Biometrics (IJCB). IEEE, pp. 1–8, 2011.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 107

Eyebrow Deserves Attention: Upper Periocular Biometrics

Hoang (Mark) Nguyen[1], Ajita Rattani[2], Reza Derakhshani[3]

Abstract: Ocular biometrics is attracting exceeding attention from research community and industry alike thanks to its accuracy, security, and ease of use in mobile devices, especially in the presence of occlusions such as masks worn during the COVID-19 pandemic. When considering the extended periocular region, eyebrows have not been getting enough attention due to their perceived low uniqueness. In this paper, we evaluate a mobile-friendly deep-learning model for eyebrow-based user authentication. Specifically, we used a fine-tuned lightCNN model for eyebrow based user authentication with promising results on a particularly challenging dataset and evaluation protocol (open-set with simulated twins). The methods achieved 0.99 AUC and 4.3% EER in VISOB dataset and 0.98 AUC and 5.6% EER on SiW datasets using closed-set and open-set analysis, respectively.

Keywords: Ocular biometrics, eyebrow biometrics, biometric recognition.

1 Introduction

Advances in deep learning has brought about remarkable improvement in the accuracy and robustness of biometric systems [Su14, PVZ15, Ng17]. Biometric systems scan a trait or modality such as face, finger or ocular region of interest in order to identify the user requesting physical or digital access. Among ocular modalities, periocular and iris have received much attention due to their accuracy and added security especially when used in smartphones [Zh18, RDR19, RD17a]. Despite advances in face recognition, there are pressing applications calling for ocular biometrics, such as users wearing face masks for safety reasons due to the recent COVID-19 pandemic. The non-touch nature of ocular biometrics adds to its utility for the aforesaid use case. However, studies have also revealed challenges related to iris and periocular recognition, including occlusions and image artifacts due to eyelids and cosmetic contact lenses, glasses, eye movements, and makeup [Bh10, MRD19, RD17b, RRD20].

Expanding the periocular region, especially towards the upper region, one may consider eyebrows and their possible utility as a biometric modality. Eyebrows, as a biometric trait, have not been well studied despite several prior works indicating their potential [Zh18, MRD19, JXS11]. Eyebrows may be used to supplement other ocular modalities such as iris in cases when the eye is closed or off-axis (Figure 1). Furthermore, eyebrow recognition can be achieved in RGB using the ubiquitous front-facing (selfie) mobile cameras,

[1] Department of Computer Science and Electrical Engineering, University of Missouri at Kansas City, hdnf39@mail.umk.edu
[2] Department of Electrical Engineering and Computer Science, Wichita State University, ajita.rattani@wichita.edu
[3] Department of Computer Science and Electrical Engineering, University of Missouri at Kansas City, derakhshanir@umkc.edu

eschewing the need for dedicated near-infrared cameras and illumination necessary for iris recognition. Due to its lower uniqueness, eyebrows are usually categorized as a soft-biometric trait [Da11]. However, thanks to their texture and morphology consistency, at least for short term mobile use cases, eyebrows maybe used for continuous user authentication or re-identification [MRD19, JXS11].

Moving from modalities to processing methods, deep learning based methods have brought about significant improvement in ocular recognition. However, many prior works in [Al18] use large neural network models, such as VGG-16 [SZ14] and ResNet [He16]. Despite advancements in mobile hardware technology, especially in inference speed, it is prudent to use models with smaller computational footprint for lower CPU and battery usage (especially for high frequency applications), faster real-time operations, and smaller download size. In this work, we employ lightCNN, a light weight convolutional neural network which uses Max-Feature-Map activation to suppress the feature map output after every convolutional layer in order to obtain compact (256-dimensional) but yet robust and accurate feature vectors for eyebrow recognition.

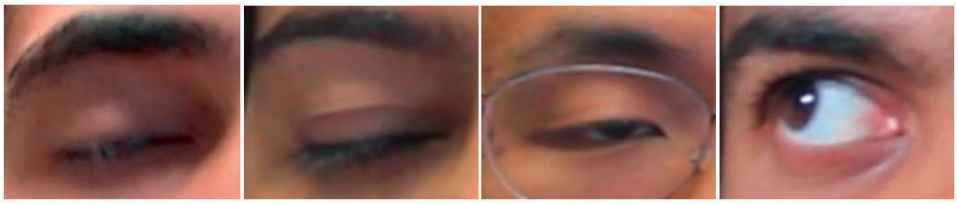


Fig. 1: Scenarios where eyebrow maybe preferable over iris for user authentication.

The aim of this work is to demonstrate capabilities of an efficient mobile eyebrow-based recognition system utilizing a single eyebrow as input for user verification under a challenging protocol including near identical eyebrows (simulated twins) and open-set evaluation. The three main contributions of this work are:

1. Establishing the utility of eyebrows as a stand-alone biometric for human recognition using smartphones' front facing cameras in presence of very challenging samples.
2. Fast and efficient end-to-end eyebrow based deep learning system including an efficient feature extraction using a light-weight CNN.
3. A thorough evaluation of the aforementioned system using open and closed set protocols on VISOB [Ra16] and SiW [LJL18] datasets, captured under different lighting conditions, along with simulated twins.

2 Prior work

The study by Xu et al. [JXS11] was the first to compare eyebrows to face and ocular recognition over a large dataset. The comparison was performed between face, eye-band, and

full eyebrow band. The authors evaluated the performance of full eyebrow band which is approximately 1/6 of the full face area using FRGC database under controlled and uncontrolled illumination settings. The study used three variants of Local Binary Patterns (LBP) for feature extraction followed by Principal Component Analysis (PCA) for dimensionality reduction. The average rank-one identification rate of the eyebrow was 31.7%.

Le et al. [LPS14] proposed an eyebrow segmentation and shape structure matching method. They used a Local Eyebrow Active Shape Model which locates 64 landmark points on the eyebrow. The model achieved 99.4% F-measure on NIST Multiple Biometric Grand Challenge (MBGC) dataset which consists of 200 images from 50 participants. For the identification task, the authors used two shape descriptors, inter-subject structure dissimilarities and intra-subject asymmetry dissimilarities, to match subjects' eyebrows. They reportedly achieved a rank one identification rate of 85.0% on a small subset and 71.3% on a large subset of the MBGC dataset.

Mohammad et al. [MRD19] investigated short-term eyebrow recognition in the presence of eyeglasses using VISOB and FERET dataset. For the short term identification using eyebrows, the authors proposed the fusion of GIST, histogram of oriented gradients (HOG), and VGG-16 features. A Support Vector Machine (SVM) classifier was used for matching. The best reported performance was 0.63% Equal Error Rate (EER) and 0.99 AUC using the fusion of the aforesaid three feature descriptors of both the eyebrows.

The summary of the state-of-the-art methods is shown in Table 1. It is worth noting that most of the existing methods used *closed-set protocol/ analysis*. Closed-set analysis, where the identities in the training and testing set overlap, usually result in higher accuracy because the system better adapts to the subject-specific peculiarities in the dataset. On the contrary, open-set evaluation identities between the training and testing set do not overlap. *To the best of our knowledge, there are no reported studies evaluating eyebrow recognition in an open-set environment, let alone with an added (simulated) twins-matching scenario.*

In order to be more relevant to real world applications at scale, the system needs to perform well in an open-set evaluation where identities in the test set are disjoint from those in the training set. Furthermore, we introduce simulated identical twin samples into our dataset, where the mirror image of users' right eyebrows are construed as new identities and matched against their left eyebrows, making our evaluation protocol even more challenging.

3 Proposed Method

3.1 Eyebrow Detection

The eyebrow region was segmented using Dlib [Ki09], an open source face landmark detection library. We used the Dlib version 19.18 that used histogram of oriented gradients (HOG) along with an ensemble of regression trees to detect 68 facial landmarks such as mouth and eye corners. We cropped the left and right eyebrow regions based on these

Tab. 1: Summary of the Prior Work on Eyebrow Recognition.

Ref	Method	Performance Metrics	Dataset	Result
[MRD19]	GIST, HOG, VGG-16, SVM	Verification rate	VISOB	99.72%
[JXS11]	LBP, WHT-LBP, DCT-LBP, DFT-LBP	Rank-1 identification rate	FRGC	31.7%
[LPS14]	Shape-Based Descriptors	Rank-1 identification rate	AR	76.0%
			MBGC	85.0%
[LLC13]	Fast Fourier Transform	Verification rate	BJUT	98.12%
			CFERET	89.22%
[LL07]	Hidden Markov Model	Verification rate	In-house	92.6%
[YXL13]	Sparsity Preserving Projection	Verification rate	In-house	92.5%

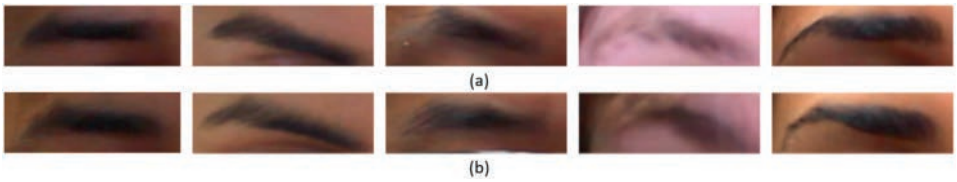


Fig. 2: Eyebrow images in SiW dataset: (a) original left eyebrow and (b) mirrored right eyebrow

landmarks. The right eyebrow crop is mirrored horizontally to synthesize a new ”twin” subject given face’s reflective symmetry, making for a challenging case similar to biometric identification of identical twins. Besides the landmarks, Dlib also provides a bounding box around the detected face.

3.2 Feature Extraction

We used lightCNN [WHS15] which has been widely used for face recognition. The general architecture of lightCNN is shown in Figure 3. The model heavily applies Max-Feature-Map (MFM) operation (see equation 1) instead of ReLu activation. This acts as feature filter after each convolution layer. The operation takes two feature maps, eliminates the element-wise minimum, and returns element-wise maximum. By doing so across feature channels, only 50% of the information-bearing nodes from each layer reach the next. Consequently, each layer is forced to preserve compact feature maps during training. The general architecture is shown in Figure 3. During the training on VISOB dataset, we added a softmax layer for classification. This layer was then removed and the remaining 256 dimensional output in MFM_fc1 was used as the feature vector representing the input identity. Two versions of lightCNN were used in this work: a 9-layer and a 29-layer lightCNN. The details of the two models can be found in [WHS15]. Thanks to their low dimensional outputs and small computational footprint for inference, both the models are suitable for mobile deployments.

$$\widehat{x}_{ij}^{k} = max(x_{ij}^{k}, x_{ij}^{N}) \tag{1}$$

3.3 Matching

Cosine similarity is used extensively in deep-learning based biometric matchers such as face recognition systems. As such, we used this metric to generate eyebrow match scores between enrollment-verification feature vector pairs obtained from our lightCNN models. The function is given below:

$$d_{cos}(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

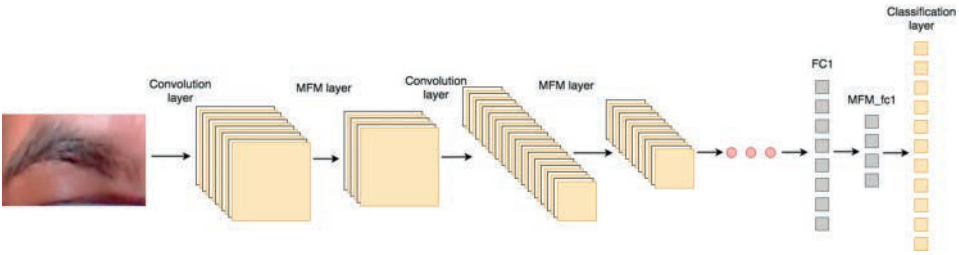


Fig. 3: Architecture of the lightCNN model used in this study.

4 Experimental Evaluation

4.1 Data and Experimental Protocol

Here we used VISible light mobile Ocular Biometric (VISOB) [Ra16] and Spoofs in the Wild (SiW) [LJL18] face anti-spoofing database to evaluate our models.

VISOB Database This database consists of eye images of about 550 healthy adults captured using three different mobile phones in three different lighting conditions. The three smartphones used in data collection are: OPPO N1, iPhone 5s, and Galaxy Note 4. During the data collection, the volunteers were asked to take selfie-like images during two visits (Visit 1 and Visit 2), 2-4 weeks apart. During each visit, images were captured in two sessions 10-15 minutes apart, and under three illumination conditions: regular office light, dim indoors, and natural daylight. In this experiment, we only used the images from OPPO device under office and natural lighting conditions.

SiW Database SiW consists of up to 8 live and 20 spoof videos from 165 participants collected at various distances, poses, illuminations, and with different facial expressions. In our experiment, we only used live videos to harvest frames. We generated more than 100,000 images from live videos by extracting one still frame from every 10 consecutive video frames. We chose the SiW dataset for our experiment because of two reasons: the rather large number of participants and the variations in eyebrow resolution. Based on the size of the detected faces' bounding boxes as delivered by Dlib, we divided the dataset into low and high resolution subsets. An eyebrow was deemed as high resolution if the pixel count in the corresponding face bounding box was larger than 200k, and considered as low resolution if such pixel count was in the 50k to 80k range.

Enrollment and Verification Data: We arranged for a total of 7 different experiments with different enrollment and verification data divisions shown in Table 2. To maintain consistency between comparisons, a single model (trained on VISOB visit 1, session 1, daylight) was used across all the experiments. In VISOB experiments, identities in the training set re-appear in testing set, thus it follows a closed-set protocol. However, all the experiments on the SiW dataset follow an open-set protocol (disjoint training-testing identities).

Data Processing and Experimental Protocol: During model training, single crop eyebrow input images were resized to 144×144 then randomly cropped to 128× 128 to fit the model input size while presenting translation variations (data augmentation). For image matching in validation and testing, we resized the image to 128× 128. We trained the models with the initial learning rate of 1e-3 for a maximum of 200 epochs and used the weights from the epoch that yielded the best validation loss (early stopping). The momentum and weight decay parameters were set to 0.9 and 10e-4, respectively.

Tab. 2: List of Experiments Conducted for Eyebrow Recognition Across Lighting, Image resolution and Time Lapse.

Dataset	Experiments	Enrollment	Verification
VISOB	Short term (Visit 1) (a)	Daylight, Session 1	Daylight, Session 2
	Short term (Visit 2) (b)	Daylight, Session 1	Daylight, Session 2
	Long term (c)	Daylight, Session 1, Visit 1	Daylight, Session 2, Visit 2
	Different illumination (d)	Daylight, Session 1, Visit 1	Office, Session 2, Visit 1
SiW	High vs. high (e)	High resolution	High resolution SiW
	Low vs low (f)	Low resolution	Low resolution SiW
	Low vs. high (g)	Low resolution	High resolution SiW

4.2 Experimental Protocol

In this study, we reflected the right eyebrow image across face's longitudinal median to double the number of identities in a way that makes the comparisons quite difficult. Given

face's reflective symmetry in the sagittal plane, such augmented dataset is similar to that of identical twins, a challenging case for face and eyebrow matching. Figure 2 shows examples of (a) left eyebrow images, (b) mirrored right eyebrow image processed using Dlib [Ki09]. Table 2 list the details of all the seven experiments conducted in this study. As mentioned earlier, we only used VISOB data collected in session 1 of visit 1 under natural light to train our model with 80% set for training and the remaining 20% for validation. We evaluated the trained models in various experiments. We used Equal Error Rate (EER) and Area Under the Curve (AUC) from ROC analysis to report classifier performance for each of the experiment in Table 2. The letter next to each experiment in table 2 indicates the corresponding ROC curve in the figure 4.

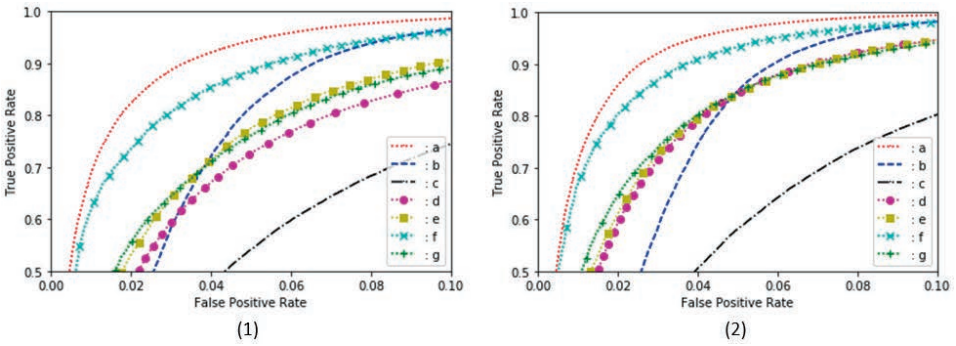


Fig. 4: ROC curves of our study's 7 experiments using (1) 9-layer and (2) 29-layer lightCNN. (a): short term verification (VISOB visit 1), (b): short term verification (VISOB visit 2), (c): long term verification, (d): different illumination, (e): high resolution vs high resolution, (f): low resolution vs low resolution, (g): low resolution vs high resolution. See Table 2 for details.

4.3 Results and Discussions

Fig 4 shows the ROC curves of the seven experiments we conducted using the 9 and 29 layer lightCNNs. As expected, both models yielded their best results on short term verification (VISOB dataset). The performance for long term verification is the worst, indicating that eyebrow is not biometrically stable over time. Cosmetic manipulation of eyebrows may also have played a role in the performance degradation. During our three experiments using SiW dataset, the low resolution versus low resolution outperformed the other two configurations. This might be due to SiW motion blur issues that are better masked in the lower resolutions.

Table 3 shows the resulting EERs [%] and AUCs (in [0,1] range). The 29-layer lightCNN yielded better results compared the 9-layer version in all the 7 experiments, meaning that the former extracted more discriminative features. Best results came from VISOB's short term verification test with the 29-layer lightCNN (EER, 4.3%, AUC, 0.990). The same network provided a 13.2% EER for VISOB long-term comparison. The 29-layer lightCNN also achieved a better EER when enrollment and verification images came from different lighting conditions (7.9% compared to 11.4% for the 9-layer model). The best open-set re-

Tab. 3: EERs and AUCs of all the Experiments in Table 2 using 9 and a 29-layer lightCNN models.

Model		LightCNN_9		LightCNN_29	
Dataset	Experiment	EER(%)	AUC	EER(%)	AUC
VISOB (Closed Set)	Short term (visit 1)	5.2	0.987	4.3	0.990
	Short term (visit 2)	7.4	0.967	6.8	0.970
	Long term	15.1	0.922	13.2	0.934
	Different illumination	11.4	0.950	7.9	0.971
SiW (Open Set)	High vs high resolution	9.7	0.963	8.0	0.973
	Low vs low resolution	7.0	0.980	5.6	0.986
	Low vs high resolution	10.3	0.960	8.2	0.973

sults (SiW dataset) show a 5.6% EER and a 0.986 AUC. Considering the especially challenging nature of our *simulated identical twins* data augmentation, these numbers show promise for eyebrows as a biometric.

One important finding from the aforementioned seven experiments is the consistency of the results across different dataset. As expected, motion blur, long term comparisons, and open set protocol did have detrimental effects on the accuracy but to a limited and reasonable extent; showing the robustness of the studied modality and matching methods.

5 Conclusion and Future Work

In this paper, we demonstrate the viability of an eyebrow recognition system that employs a light-weight deep learning model and operates on selfie-like captures. We do so using a challenging data augmentation pipeline akin to comparing identical twins, and extend our experiments to long term, open set protocols to show the resiliency of the proposed modality and matching method. Such non-touch ocular methods are especially important during challenging times such as the recent COVID-19 pandemic that has rendered ubiquitous face recognition systems into a hassle for large swaths of users wearing protective face masks. Eyebrows do deserve our attention. As a part of the future work, we would like to evaluate our pipeline with different datasets using different deep learning models in fully open-set environment. Further, eyebrow recognition will be compared with other periocular regions such as iris. Lastly, an adaptive system will be proposed to fuse eyebrow with other intra-ocular regions to further enhance the performance.

6 Acknowledgement

This work was made possible in part by a grant from ZOLOZ. Dr. Derakhshani is also a consultant for the company.

References

[Al18] Alahmadi, A. A.; Hussain, M.; Aboalsamh, H.; Zuair, M.: ConvSRC: SmartPhone based Periocular Recognition using Deep Convolutional Neural Network and Sparsity Augmented Collaborative Representation. CoRR, abs/1801.05449, 2018.

[Bh10] Bharadwaj, S.; Bhatt, H. S; Vatsa, M.; Singh, R.: Periocular biometrics: When iris recognition fails. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, pp. 1–6, 2010.

[Da11] Dantcheva, A.; Velardo, C.; D'angelo, A.; Dugelay, J.: Bag of soft biometrics for person identification. Multimedia Tools and Applications, 51(2):739–777, 2011.

[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 (CVPR). June 2016.

[JXS11] Juefei-Xu, F.; Savvides, M.: Can your eyebrows tell me who you are? In: 2011 5th International Conference on Signal Processing and Communication Systems (ICSPCS). IEEE, pp. 1–8, 2011.

[Ki09] King, Davis E.: Dlib-ml: A Machine Learning Toolkit. J. Mach. Learn. Res., 10:1755–1758, December 2009.

[LJL18] Liu, Yaojie; Jourabloo, Amin; Liu, Xiaoming: Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 2018.

[LL07] Li, Y.; Li, X.: HMM Based Eyebrow Recognition. In: Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007). volume 1, pp. 135–138, 2007.

[LLC13] Li, Y.; Li, H.; Cai, Z.: Human eyebrow recognition in the matching-recognizing framework. Computer Vision and Image Understanding, 117(2):170–181, 2013.

[LPS14] Le, T Hoang Ngan; Prabhu, Utsav; Savvides, Marios: A novel eyebrow segmentation and eyebrow shape-based identification. In: IEEE International Joint Conference on Biometrics. IEEE, pp. 1–8, 2014.

[MRD19] Mohammad, A. S.; Rattani, A.; Derakhshani, R.: Eyebrows and eyeglasses as soft biometrics using deep learning. IET Biometrics, 8(6):378–390, 2019.

[Ng17] Nguyen, K.; Fookes, C.; Ross, A.; Sridharan, S.: Iris recognition with off-the-shelf CNN features: A deep learning perspective. IEEE Access, 6:18848–18855, 2017.

[PVZ15] Parkhi, O. M; Vedaldi, A.; Zisserman, A.: Deep face recognition. In: bmvc. volume 1, p. 6, 2015.

[Ra16] Rattani, A.; Derakhshani, R.; Saripalle, S. K.; Gottemukkula, V.: ICIP 2016 competition on mobile ocular biometric recognition. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 320–324, Sept 2016.

[RD17a] Rattani, A; Derakhshani, R: Ocular biometrics in the visible spectrum: A survey. Image and Vision Computing, 59:1 – 16, 2017.

[RD17b] Rattani, A.; Derakhshani, R.: On fine-tuning convolutional neural networks for smartphone based ocular recognition. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, pp. 762–767, 2017.

[RDR19] Rattani, A.; Derakhshani, R.; Ross, A.: Selfie Biometrics. 2019.

[RRD20] Reddy, N.; Rattani, A.; Derakhshani, R.: Generalizable deep features for ocular biometrics. Image and Vision Computing, 2020.

[Su14] Sun, Y.; Chen, Y.; Wang, X.; Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. pp. 1988–1996, 2014.

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

[WHS15] Wu, X.; He, R.; Sun, Z.: A Lightened CNN for Deep Face Representation. CoRR, abs/1511.02683, 2015.

[YXL13] Yang, X.; Xu, X.; Liu, C.: Eyebrow recognition based on sparsity preserving projections. In: IEEE Conference Anthology. IEEE, pp. 1–4, 2013.

[Zh18] Zhang, Q.; Li, H.; Sun, Z.; Tan, T.: Deep Feature Fusion for Iris and Periocular Biometrics on Mobile Devices. IEEE Transactions on Information Forensics and Security, 13(11):2897–2912, Nov 2018.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 117

End-to-end Off-angle Iris Recognition Using CNN Based Iris Segmentation

Ehsaneddin Jalilian[1], Mahmut Karakaya[2], Andreas Uhl[3]

Abstract: While deep learning techniques are increasingly becoming a tool of choice for iris segmentation, yet there is no comprehensive recognition framework dedicated for off-angle iris recognition using such modules. In this work, we investigate the effect of different gaze-angles on the CNN based off-angle iris segmentations, and their recognition performance, introducing an improvement scheme to compensate for some segmentation degradations caused by the off-angle distortions. Also, we propose an off-angle parameterization algorithm to re-project the off-angle images back to frontal view. Taking benefit of these, we further investigate if: (i) improving the segmentation outputs and/or correcting the iris images before or after the segmentation, can compensate for off-angle distortions, or (ii) the generalization capability of the network can be improved, by training it on iris images of different gaze-angles. In each experimental step, segmentation accuracy and the recognition performance are evaluated, and the results are analyzed and compared.

Keywords: Off-angle iris segmentation, Off-angle iris recognition, Iris parameterization, Convolutional neural network, CNN.

1 Introduction

Iris recognition is known to be one of the most accurate biometric recognition techniques, widely adopted for many security needs in recent years. Accuracy of these systems, however, relies highly on the accurate segmentation of the iris texture in the captured eye images. Ever since the first iris recognition system proposed by John Daugman [Da09], a wide variety of techniques has been proposed to perform segmentation in eye images captured typically in a frontal view, under a controlled or constrained environment. In practice however, many of the users or operators of these systems are inexperienced and often capture images where the subjects are looking in the wrong direction due to inadvertent eye movement. Also, the emerging standoff iris biometric systems and the recent trend towards ”on-the-move-acquisition” are transforming iris biometric systems from being operated in well-controlled setup, to being smart standoff modalities. The iris images captured under such conditions are more likely to be off-angle, and incorporate additional off-angle related distortions.

[1] Department of Computer Science, University of Salzburg, Jakob-Haringer Str.2, Salzburg, Austria ejalilian@cs.sbg.ac.at
[2] Department of Computer Science, Kennesaw State University, 1100 South Marietta Pkwy.Marietta, GA, USA mkarakay@kennesaw.edu
[3] Department of Computer Science, University of Salzburg, Jakob-Haringer Str.2, Salzburg, Austria Uhl@cs.sbg.ac.at

Segmentation tasks in such images become quite challenging as the iris boundaries are dilated, of elliptical shape, or even missing in the extreme off-angle images. Most classical segmentation approaches which are mainly based on the integro-differential, circular Hough Transform, and edge detection techniques, which rely on visibility of clear iris contours, fail to perform segmentation in such images. Consequently, also most feature comparison algorithms operating under the assumption that the iris texture lies on a flat frontal plane and possesses a circular geometric property, fail to perform the comparison task properly as well [ZA10]. Addressing such challenges, off-angle iris recognition has became a hot research topic within the biometrics community recently.

With recent advancement in deep learning techniques, some convolutional neural networks (CNN) were proposed for the challenging task of iris segmentation (*e.g.* [Ar18] [JU17]). While the proposed models proved to preform superior to the classical segmentation methods, yet the scarce researches dedicated to parameterization and normalization of obtained iris segmentations are just limited to frontal iris images, and no comprehensive recognition framework has been introduced for off-angle iris recognition using such modules. Jalilian *et al.* [JUK19] studied the effect of off-angle distortions on the segmentation performance of CNNs. We extend this study by specifically investigating the effect of different gaze-angles on the subsequent recognition performance. First, as a distinction to the segmentation studies in [JUK19], here we introduce a segmentation improvement scheme to compensate for some degradations in the segmentation masks, caused by the off-angle distortions. In this framework, we propose an off-angle parameterization method to determine the extent of off-angle-ness and to geometrically re-project the segmentations and their corresponding off-angle iris images back to frontal view. We further define several variants of end-to-end recognition pipelines to enable the usage of the CNN based segmentations for the final task of recognition. In the first approach, termed ”improved-homogeneous”, we train a dedicated CNN with homogeneous iris images of each distinct gaze-angle, and then carry out segmentation in iris images with certain gaze-angles. The segmentation outputs then are improved, and both the segmentation and recognition performance are evaluated afterwards. In the second approach, denoted as ”improved-heterogeneous”, we propose a heterogeneous-angle training, in which a network trained with iris images exhibiting different gaze-angles, is applied to iris images with any gaze-angle. Here we target to improve the generalization capability of the networks used in the improved-homogeneous approach, in a way that we can obtain hopefully better results than we obtained using the angle-specific training configuration. In the third approach we utilize our off-angle parameterization method (as explained in Section 3) to geometrically re-project the corresponding off-angle iris images back to frontal view before applying unwrapping and normalization. We denote this approach as ”corrected-homogeneous.” Doing so, we hope to correct the off-angle iris texture, compensating for the degradations imposed by the off-angle distortions, and thus enhance the biometric data encoded into it. And finally, by analogy to the corrected-homogeneous approach, we considered the ”corrected-heterogeneous” approach, in which we investigate the effect of the correction mechanism on the recognition performance using a heterogeneous training configuration.

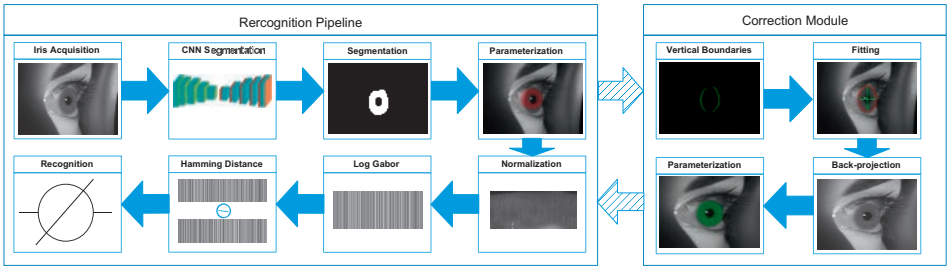


Figure 1: Recognition pipeline and the correction module

2 Related Work

Several different techniques have been proposed to address the off-angle iris segmentation and recognition problem. For example Daugman proposed to detect inner and outer off-angle iris boundaries using an active contour method, based on the discrete Fourier series expansion of the contour data [Da06]. Shah and Ross combined snakes segmentation with geometric active contours [SR09]. Zuo *et al.* [ZS09] used intensity, shape, and localization features from the iris and pupil to automatically segment non-ideal iris images. Their method demonstrated performance improvement on challenging iris images up to 30°. Price *et al.* [R-07] developed a generalized eye model to correct for perspective and refractive distortion of the iris pattern using ray tracing techniques. They reported a median reduction of Hamming Distance for synthetic eyes with gaze up to 60°. Recent advances in deep learning techniques enabled the application of deep neural networks for iris segmentation. For example Liu *et al.* [Li16b] proposed two iris segmentation techniques based on different topologies of CNNs (hierarchical convolutional neural networks and multi-scale fully convolutional networks). The method presented by Arsalan *et al.* [Ar17] roughly estimates the iris region using an edge detection algorithm and then classifies the pixels in two classes (iris and non-iris) by using a CNN. The study presented in [JU17] utilized a fully convolutional encoder-decoder network trained for classifying iris and non-iris pixels in images acquired in a wide set of heterogeneous conditions, including off-angle images. The work presented in [Ar18] proposed a deep network called IrisDenseNet, which is based on VGG-16, to deal with low quality iris images, such as side views, glasses, off-angle eye images and rotated eyes. There are far more approaches dedicated for eff-angle iris segmentation/recognition. Yet due to the space limitation, we narrowed our review to the methods presented above. To review further approaches please refer to *e.g.* [S-16].

3 Off-angle Iris Parameterization and Segmentation Improvement

Off-angle iris parameterization: The available algorithms used for parameterization of the iris region in the CNN based segmentation are limited to the frontal segmentation outputs, where circular Hough transform is used to parameterize the iris region. The main obstacle to apply an elliptic parameterization (as the iris shape looks in the off-angle view)

is the tendency of such models to overly oblong or obround, due to occlusion of the iris by eyelids or eyelashes. To resolve this issue, we propose to search only for the vertical edges in the segmentation outputs. The resulting edge points secure the proper fitting of an ellipse to the actual iris region (see Figure 1 for an example). In the next step, we extract the horizontal and vertical axes information of the ellipse, and use them for re-projecting (correcting) the segmentation outputs and their corresponding off-angle iris images back to frontal view as follows. Assuming that our ellipse is in the following parametric form:

$$x = x_0 + Q \times \begin{bmatrix} a \times cos(\theta) \\ b \times sin(\theta) \end{bmatrix}, \tag{1}$$

where x and x_0 are 2-dimensional vectors, and $a > b > 0$ correspond to the horizontal and vertical axes of the ellipse, respectively. Q is the rotation matrix, and θ represent the rotation angle. We assume a vertical ellipse, Thus:

$$Q = \begin{bmatrix} cos(90) & -sin(90) \\ sin(90) & cos(90) \end{bmatrix}. \tag{2}$$

We want our transformation to produce y in the shifted, rotated coordinates:

$$y = \begin{bmatrix} 1 & 0 \\ 0 & a/b \end{bmatrix} \begin{bmatrix} a \times cos(\theta) \\ b \times sin(\theta) \end{bmatrix}, \tag{3}$$

and x in the original coordinates. Submitting to the equation (1), we can infer the affine transformation matrix we need to re-project the parameterized ellipse back to frontal view, so that it possess circular shape:

$$x = \left[Q \begin{bmatrix} 1 & 0 \\ 0 & a/b \end{bmatrix} Q' \right] x + \left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - Q \begin{bmatrix} 1 & 0 \\ 0 & a/b \end{bmatrix} Q' \right] x_0. \tag{4}$$

Segmentation Improvement: We improved the segmentation outputs by applying some morphological operations. It was already understood that the network tends to produce some false-positive detection, in specific, along the segmentation output masks borders [JUK19]. So, we first defined a marginal area (A) along each border of the segmentation output masks (with a width (in pixel) equal to $1/5$ of the length of the same border), and then performed an opening operation with a big (disk-shape) structuring element (B):

$$A \circ B = (A \ominus B) \oplus B, \tag{5}$$

Figure 2: Sample iris image with P0 gaze-angle and its corresponding segmentation (green-color) and error mask (red-color) before (middle), and after correction (right), using the network trained on P0 images

where $\ominus$ and $\oplus$ denote erosion and dilation, respectively. We further performed another opening operation on the whole segmentation outputs using a small (disk-shape) structuring element to remove small false-positive detections outside the iris region. Figure 2 shows a sample segmentation output and its corresponding improved segmentation mask.

4 Experimental Framework

Dataset: For our experiments we used a subset (containing 4400 left eye iris images captured from 40 subjects) of an off-angle iris database [Ka13]. The iris images in this database are captured by two near-infrared sensitive IDS-UI-3240ML-NIR cameras. Images at 0°gaze-angle were captured by a frontal fixed camera, and off-angle images were captured by a frontal moving camera rotating horizontally from -50°(N50) to +50°(P50) in angle with a 10°step-size. Each camera captured 10 iris images per stop, giving 10 frontal and 100 off-angle iris images captured from each subject, to comprise 400 images per angle (examples of images in the database are presented in Figure 3). The database is accessible on request (from the authors), and further details about it can be found in [Ka13]. We developed the ground-truth labels (required for training the network) for all images available in the dataset using the iris, pupil, upper and lower eyelid parameters specified manually. For our experiments we divided the whole dataset into two equal parts (each containing iris images of 20 separate subjects), and used one part as our testing data and the other one as our training data.

Fully convolution neural network (FCN): We selected the RefineNet [Li16a] to perform the iris segmentations in our experiments. The network is already proven to enable high-resolution prediction, and at the same time, preserve the boundary information (which is needed for our parameterization mechanism). The network is a multi-resolution refinement network, which employs a 4-cascaded architecture with 4 Refining units, each of which directly connects to the output of one Residual net [He15] block, as well as to the preceding Refining block in the cascade. Each Refining unit consists of two residual convolution units (RCU), which include two alternative ReLU and 3×3 convolutional layers. The output of the RCU units are processed by 3×3 convolution and up-sampling layers incorporated in multi-resolution fusion blocks. A chain of multiple pooling blocks, each consisting a 5×5 max-pooling layer and a 3×3 convolution layer, next operate on the feature maps, so that one pooling block takes the output of the previous pooling block as input. Therefore, the current pooling block is able to re-use the result from the previous

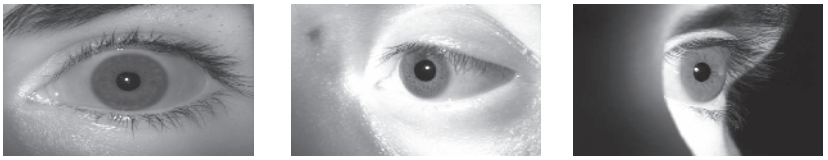


Figure 3: Sample iris images with P0 (left), N50 (middle), and P50 (right) gaze-angles

pooling operation and thus access the features from a large region without using a large pooling window. Finally, the outputs of all pooling blocks are fused together with the input feature maps through summation of residual connections. We used ADAM optimizer with learning rate of 0.0001, executing 40,000 iteration to train the network. The implementation of the network was realized in Keras using TensorFlow back-end.

Recognition Pipeline: The output segmentations (after applying correction or improvement), are parameterized using the technique introduced in [HJU19]. The extracted iris patterns are normalized by unwrapping the circular region into a rectangular block of constant dimensions. The algorithm repeats the last pixel for a given angle if no values are available. Each isolated iris pattern is then demodulated to extract its phase information (feature) using quadrature 1-D Gabor wavelets. To compare the unique extracted features to each other, the Hamming distance with rotation correction were calculated in the comparison phase. We used the University of Salzburg implementation of these algorithms, as provided in the Iris Toolkit (USIT)[3]. Figure 1 illustrates the overall recognition pipeline, along with the proposed parameterization and correction module.

Segmentation Evaluation and Measures: In order to facilitate proper quantification of the accuracy of the segmentations in each experiment, we considered the $nice1$ iris segmentation error rate, which is based on the NICE1 protocol[4], as used in several iris segmentation challenges. Accordingly, the segmentation error rate $(nice1)$ for each segmentation output mask I_i is given by the proportion of corresponding disagreeing pixels (through the logical exclusive-or operator) with the ground-truth mask, over all the output mask as:

$$nice1 = \frac{1}{c \times r} \sum_{c'} \sum_{r'} O(c', r') \otimes C(c', r'), \tag{6}$$

where c and r are the dimensions of the segmentation, and $O(c', r')$ and $C(c', r')$ are, respectively, pixels of the segmentation and the ground-truth mask. The value of $(nice1)$ is in the [0, 1] interval, and 1 and 0 are the worst and the best scores, respectively.

[3] http://www.wavelab.at/sources/USIT
[4] http://nice1.di.ubi.pt/

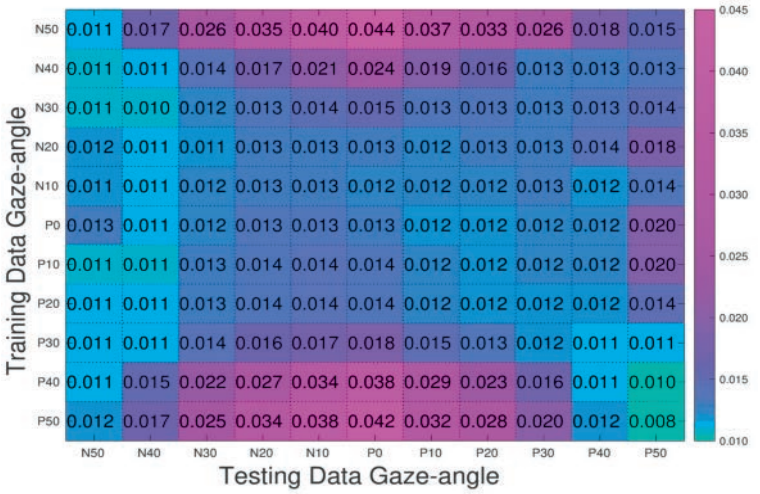


Figure 4: Segmentation performance using the improved-homogeneous approach

5 Experiments and Analysis

We initiated our experiments by investigating the effect of different gaze-angles on the CNN based off-angle iris segmentations, after the improvement, as well as evaluating their subsequent recognition performance (under the improved-homogeneous approach). To focus our experiments on this objective, we considered an ideal (but unrealistic) condition, in which the true images' gaze-angles are already known. "Theoretically," one may use the horizontal and vertical axes information to estimate the images gaze-angles (D) using: $D = acosinus(HorizontalAxis/VertcialAxis)$. So, we trained a dedicated network with iris images belonging to each distinct gaze-angle separately, and then performed segmentation in all our testing data, and improved the segmenation outputs as already described in Section 3. Figure 4 shows the results, as average $nice1$ error for this experiment. Affirming to what we found using the identical training scheme (Homogeneous) and network (RefineNet) already in [JUK19], we can see the direct relation of the network performance to the similarity of gaze-angles of the training and testing images, here after the morphological improvement too. Yet the key new finding is that, the performance gradually improves as the gaze-angles of the training and testing data converge in terms of angle but may also diverge in terms of the direction. To be more precise, the network is able to detect the symmetric iris elliptical features in the images captured from the same angle (with respect to frontal view), but in opposite direction. The applied improvement, which in fact compensated for some false-positive detections (caused by the off-angle distortions), allowed us to figure out this capability of the network. Overall, the applied improvement resulted in considerable enhancements in almost all segmentation results (especially for the right off-angle (P) images), compared to the segmentation results obtained in [JUK19], as the average error decreased (about 47%) from 0.030 to 0.016.

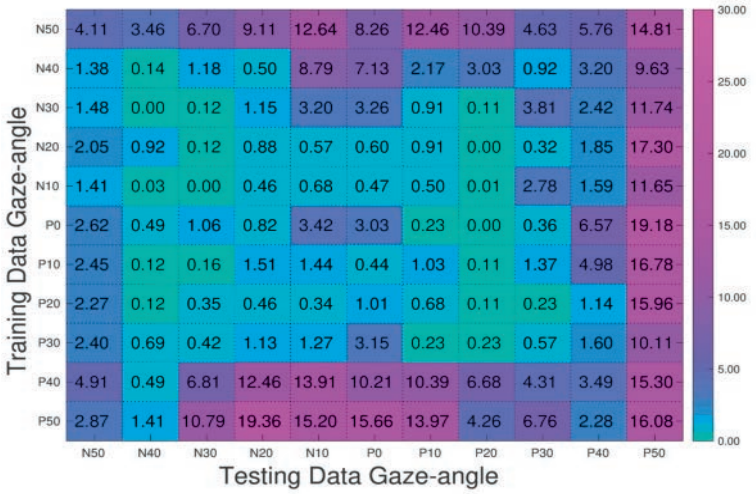


Figure 5: Recognition performance using the improved-homogeneous approach

In the next step, we fed the improved segmentations along with their corresponding images to the recognition pipeline to investigate the recognition performance in terms of *EER*. Figure 5 shows the results for this experiment. Expectedly, we can observe that the segmentation results are translated into the recognition scores, following the same trends already discussed in the segmentation experiments. The only visible difference here is the lower recognition performance of the extreme gaze-angle images (*i.e.* N50 and specially P50). This seems mainly to be due to the extreme 3D and perspective erosion of the extracted iris texture, which leads to the lower recognition performance on these images. In the improved-heterogeneous approach, we considered to investigate if we can improve the generalizability of the network by switching to a heterogeneous training setting, where we include iris images with different gaze-angles into the training data. We tested the trained network in all iris images in our testing data, applied the improvement, and evaluated performance afterward, differentiating and grouping results into the different gaze-angles available. While the heterogeneous configuration was expected to deliver good results (compared to the angle-specific training configuration), based on the findings in [JUK19], here we (i) evaluated the extent to which the improvement applied can enhance the segmentation performance, and (ii) verified if the improved segmentations can eventually improve the recognition performance, beyond the improved angle-specific training configuration. Figure 6 demonstrates the segmentation results for this experiment in the form of Boxplot for each gaze-angle group (after the improvement). As the results show, applying the improvement, we obtained a considerable enhancement in almost all segmentation results (especially for the right off-angle (P) images), compared to the angle-specific improved-homogeneous results already obtained, as well as those obtained in the identical heterogeneous configuration without improvement in [JUK19], as the average segmentation error decreased (about 4.5 times) from 0.023 to 0.005. Figure 7 shows

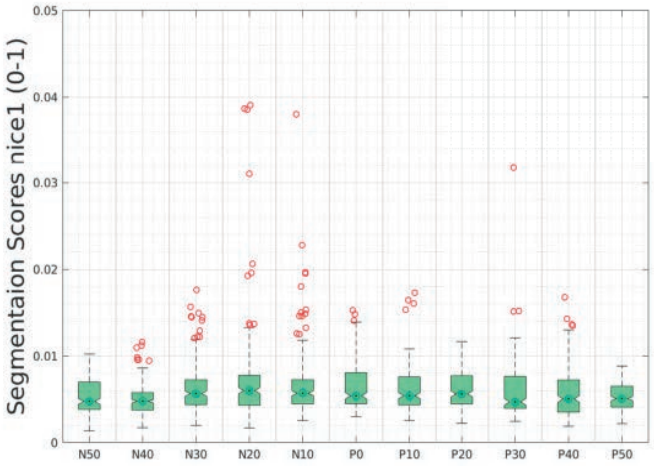


Figure 6: Segmentation performance using the improved-heterogeneous approach

the subsequent recognition results obtained using the corresponding images. Excluding a slight declination in the recognition results of N40 gaze-angle images, all other results show considerable improvements, compared to the angle-specific configuration results (the improved-homogeneous) shown in Figure 5. Of course, this is a positive result, as it enables us to refrain from the angle-specific training strategy, and even better, there is no need to determine the iris images gaze-angles or carry out the correction.

In the corrected-homogeneous approach, we target to address if re-projecting the off-angle iris images back to frontal view and correcting the off-angle iris texture can compensate for the degradations imposed by the off-angle distortions, and eventually improve the system recognition performance. To address this, we first applied our parameterization algorithm (already explained in Section 3) to the improved segmentation outputs obtained in the previous step, and subsequently re-projected them along with their corresponding iris images back to frontal view. The corrected data then was fed into the recognition pipeline to evaluate the recognition performance. Figure 8 shows the recognition results for this experiment. When comparing the results to those obtained using the improved-homogeneous approach, we can only observe slight improvements in the results of configurations where the training and testing data are close to frontal (*i.e.* P0, P10, P20, ...) view, as well as the extreme gaze-angles (*i.e.* , P50 and N50), where the gaze-angles of the training and teasing data are the same. For the rest of configurations, the results gradually degrade as we move towards the right and, in specific, the left sides of the table (compared to the

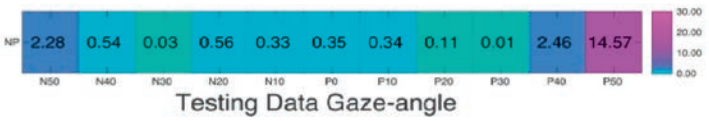


Figure 7: Recognition performance using the improved-heterogeneous approach

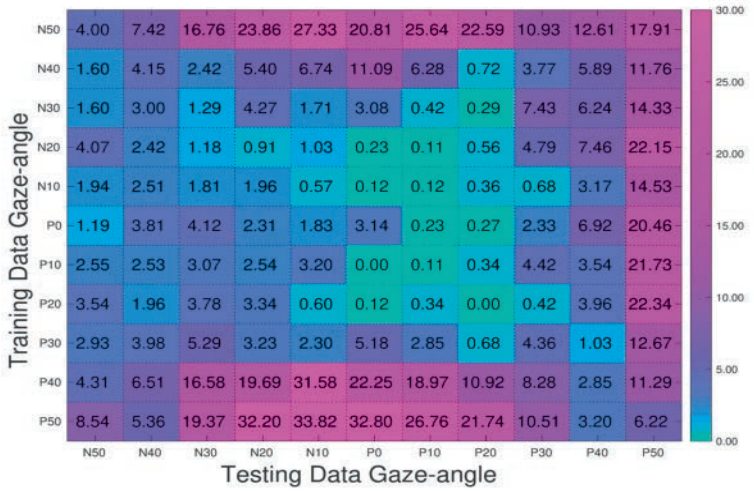


Figure 8: Recognition performance using the corrected-homogeneous approach

corresponding results, obtained using the improved-homogeneous approach, presented in the table in Figure 5. We can infer two degradation factors analyzing these results. First, the interpolation applied during the correction procedure starts to agonize the biometric features encoded in the iris texture, as the images' gaze-angle get far from frontal view, and the amount of the interpolation applied increases. Second, possible imperfections of the correction algorithm, may result in some differences in iris images belonging to each distinct subject, which eventually lead to degradation of genuine scores and subsequent recognition performance of the system. The pattern and scale of the changes in the results are a function of influence of these two factors.

We further considered the corrected-heterogeneous approach, in which we investigated if correcting the off-angle iris texture can compensate for the degradations imposed by the off-angle distortions, and thus improve the recognition performance, within a heterogeneous training configuration. So here, after training the network on iris images with different gaze-angles, and testing it on the images of each gaze-angle separately, the segmentation outputs were morphologically improved, parameterized and re-projected back to frontal view, and the recognition performance was evaluated subsequently. Figure 9 demonstrates the results for this experiment per gaze-angle. As it can be seen in the figure, the performance pattern is similar to what we found already in the corrected-homogeneous

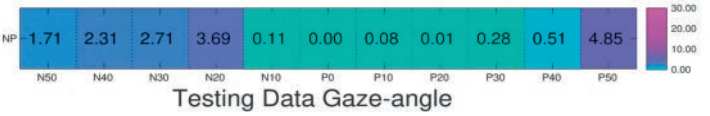


Figure 9: Recognition performance using the corrected-heterogeneous approach

approach. To be more precise, while we can see considerable improvements in the overall recognition results (compared to the corresponding results obtained using the angle-specific corrected-homogeneous approach) due to the supremacy of the heterogeneous configuration used, yet the same performance degradations (*i.e.* in the results of N40, N30, P30 gaze-angles) and enhancements (*i.e.* in the results of N50, P0, P10, P50 gaze-angles), as observed in the corrected-homogeneous approach, are visible here too.

6 Conclusion

The morphological improvement technique proved to compensate for some off-angle related segmentation degradations, enhancing the segmentation and the recognition results beyond those obtained in [JUK19], in identical configurations. The experiments carried out under the improved-homogeneous approach showed that the network performance gradually improves as the gaze-angle of the training and testing data converges in terms of angle but diverges in terms of direction. This showed the capability of the network to detect the symmetric iris contents in the images captured from the same angle, but in the opposite direction, which was figured out as the result of the segmentation improvement done. The experimental results of the viewing angle correction based approaches showed that the interpolation applied during the correction procedure and the possible imperfections of the correction algorithm, can dominantly influence the distinction of the iris images and thus undermine their subsequent recognition performance. This leads us to the conclusion that: Unless applying it to iris images with closed-to-frontal gaze-angles (*i.e.* up to 20°), and performing perfect (error free) correction, this angle correction based approaches are not expected to deliver promising recognition results (specially on the +20°off-angle images), when applied on the CNN based off-angle segmentations. While the heterogeneous training approaches were already expected to deliver good results (compared to the angle-specific homogeneous training configurations), based on the findings in [JUK19], yet our experiments actually showed that the applied segmentation improvement enhances the segmentation results, beyond those obtained using the same configuration (heterogeneous) in [JUK19], as well as improving the recognition results beyond the angle-specific training configuration results. In practice, this was very positive result, as it enabled us to refrain from the angle-specific training strategy, and even from the need for correcting the images' gaze-angles before being able to deploy the recognition systems.

Acknowledgment

This project was partly funded from the FFG KIRAS project AUTFingerATM under grant No. 864785 and the FWF project "Advanced Methods and Applications for Fingervein Recognition" under grant No. P 32201-NBL.

References

[Ar17] Arsalan, Muhammad; Gil-Hong, Hyung; Ali-Naqvi, Rizwan; Beom-Lee, Min; Cheol-Kim, Min; Seop-Kim, Dong; Sik-Kim, Chan; Ryoung-Park, Kang: Deep learning-based

iris segmentation for iris recognition in visible light environment. Symmetry, 9(11):263, 2017.

[Ar18] Arsalan, Muhammad; Ali-Naqvi, Rizwan; Seop-Kim, Dong; Ha-Nguyen, Phong; Owais, Muhammad; Ryoung-Park, Kang: IrisDenseNet: Robust iris segmentation using densely connected fully convolutional networks in the images by visible light and near-infrared light camera sensors. Sensors, 18(5):1501, 2018.

[Da06] Daugman, John: Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. Proceedings of the IEEE, 94(11):1927–1935, 2006.

[Da09] Daugman, John: How iris recognition works. In: The essential guide to image processing, pp. 715–739. Elsevier, 2009.

[He15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep Residual Learning for Image Recognition. CoRR, abs/1512.03385, 2015.

[HJU19] Hofbauer, Heinz; Jalilian, Ehsaneddin; Uhl, Andreas: Exploiting superior CNN-based iris segmentation for better recognition accuracy. Pattern Recognition Letters, 120:17–23, 2019.

[JU17] Jalilian, Ehsaneddin; Uhl, Andreas: Iris Segmentation Using Fully Convolutional Encoder–Decoder Networks. In (Bir Bhanu, Ajay Kumar, ed.): Deep Learning for Biometrics, chapter 6, pp. 133–155. Springer, (ZG) Switzerland, 2017.

[JUK19] Jalilian, Ehsaneddin; Uhl, Andreas; Karakaya, Mahmut: Gaze-angle Impact on Iris Segmentation using CNNs. In: Proceedings of the IEEE 10th International Conference on Biometrics: Theory, Applications and Systems. Tampa, Florida, USA, pp. 1–8, 2019.

[Ka13] Karakaya, Mahmut; Barstow, Del; Santos-Villalobos, Hector; Thompson, Joseph: Limbus impact on off-angle iris degradation. In: International Conference on Biometrics (ICB). pp. 1–6, 2013.

[Li16a] Lin, Guosheng; Milan, Anton; Shen, Chunhua; D-Reid, Ian: RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation. CoRR, abs/1611.06612, 2016.

[Li16b] Liu, Nianfeng; Li, Haiqing; Zhang, Man; Liu, Jing; Sun, Zhenan; Tan, Tieniu: Accurate iris segmentation in non-cooperative environments using fully convolutional networks. In: 2016 International Conference on Biometrics (ICB). IEEE, pp. 1–8, 2016.

[R-07] R-Price, Jeffery; F-Gee, Timothy; Paquit, Vincent; W-Tobin, Kenneth: On the efficacy of correcting for refractive effects in iris recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–6, 2007.

[S-16] S-Bolme, David; ; Santos-Villalobos, Hector; Thompson, Joseph; Karakaya, Mahmut; Boehnen, Chris Bensing: Off-Angle Iris Correction Methods. In: Handbook of Iris Recognition. Springer London, London, pp. 497–518, 2016.

[SR09] Shah, Samir; Ross, Arun: Iris segmentation using geodesic active contours. IEEE Transactions on Information Forensics and Security, 4(4):824–836, 2009.

[ZA10] Zuo, Jinyu; A.Schmid, Natalia: On a Methodology for Robust Segmentation of Nonideal Iris Images. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40(3):703–718, June 2010.

[ZS09] Zuo, Jinyu; Schmid, Natalia: On a methodology for robust segmentation of nonideal iris images. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40(3):703–718, 2009.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 129

Iris Recognition in Postmortem Enucleated Eyes

Sashi K. Saripalle [1], Adam McLaughlin[2], Reza Derakhshani[3]

Abstract: This paper presents a comprehensive multispectral study of iris recognition on postmortem enucleated eyes over a period of three days. An off the shelf iris recognition methodology is employed to analyze the biometric capability of iris in the post mortem setting. We observed that iris patterns of enucleated eyes can provide biometric matches with no false accepts for up to 164 hours after death, albeit with high false rejection rates. We also present our observations on the effects of the environment and other confounding factors that may affect the performance of postmortem iris recognition, with recommendations for rehydration of specimen to regain postmortem biometric utility.

Keywords: Postmortem biometrics, iris recognition, biometrics, forensics.

1 Introduction

Human iris patterns are arguably among the best biometric modalities for personal identification due to their unique and stable textures. It has been observed that the matching probability of two different irides is about one in seven billion [JD01]. However, when it comes to postmortem settings, the functionality of iris as a biometric requires further investigations. Previous work from Warsaw University details several aspects of postmortem iris recognition, confirming that modern iris recognition systems can indeed match postmortem iris samples [MAP18, MAP16a, MAP16b]. However, many aspects of postmortem iris biometrics, especially with off the shelf iris recognition systems under different environmental and other external factors is not widely studied. In this work we revisit this challenging area of biometrics by collecting a new postmortem dataset to provide further information on the functionality of traditional iris biometric systems when applied to enucleated human eyes, and the effect of external environment factors on postmortem iris recognition.

To achieve these goals, we started our study by collecting a new dataset by capturing postmortem enucleated human eyes using a multispectral visible-IR camera and illuminator over a period of 96 hours. The choice of multispectral imaging was due to the reported postmortem iris color changes noted in pigs [EMD08]. Although not well-studied in humans, we noted the possibility of iris color changes in postmortem human iris. The second reason for multispectral imaging was to test the hypothesis that there could be faint blackbody afterglow signature pursuant to an infrared flash. This hypothesis was based on blood

[1] Computational Intelligence and Bio-Identification Technologies Lab, University of Missouri at Kansas City, ssqnf@mail.umkc.edu
[2] School of Medicine , University of Missouri – Kansas City, adammclaughlin81@gmail.com
[3] Computational Intelligence and Bio-Identification Technologies Lab, University of Missouri at Kansas City, derakhshanir@umkc.edu

spotting technique using lock-in amplifier for detection of black-body radiation of blood proteins similar to the method developed by Morgan [SM11]. However, we were not able to observe any changes immediately after infrared flash on iris tissue within the limitations of our experimental setup, and thus we won't be expanding upon it during this paper.

2 Data Collection

2.1 Hardware

The main study camera was a MS4100 multispectral unit (Optech Intl.) with multiple band-passed CCDs. The distance between the specimen and lens of the camera was 0.22 meters. We also built a custom multi-spectral ring light along its control unit to match our camera's centre multi-spectral frequencies. The light control unit also triggered camera's shutter based on a predefined sequence.

2.2 Capture Process

We obtained the study's enucleated human eyes from a local eye bank in temperature-controlled boxes stabilized at 2 degree Celsius. The eyes were then transported to a medical facility where they were stored under the supervision of an ophthalmologist. From twenty-seven acquired specimen, a total of eleven specimen were used for this study. Sixteen of the specimens were rejected due to either perforated iris tissue or punctured eyes balls at the start of study (figure 1). An ophthalmologist inspected the specimen to make sure the structure of eyeball and iris tissue were intact before the study was initiated on the specimen. The average time from death to procurement of eyes was around 7.5 hours, with the minimum of and maximum time lapses being 3.5 and 24 hours, respectively. The average time of procurement to study was six days. The average age of donor was 72 years. Analysis of our specimen was done in two parts: short-term and long-term. A total of 214 samples were used for short term analysis, and a total of 718 samples were used for long term analysis. The eye colors of the specimen were either light brown, blue, green, gray, or dark brown. Further experimental setup details can be found in section V.

We also captured data from fifty live control subjects over two sessions at UMKC under the auspices of an Institutional Review Board approved protocol. These sessions were at least one week apart, and multiple captures were acquired for each session with a time delay of at least 15 minutes between captures. The iris scores from control subjects were used to set the threshold for postmortem analysis.

3 Related Work

Postmortem iris recognition has been receiving less attention in literature in part due to a belief that iris possesses little biometric value after death. Other barriers to the study of

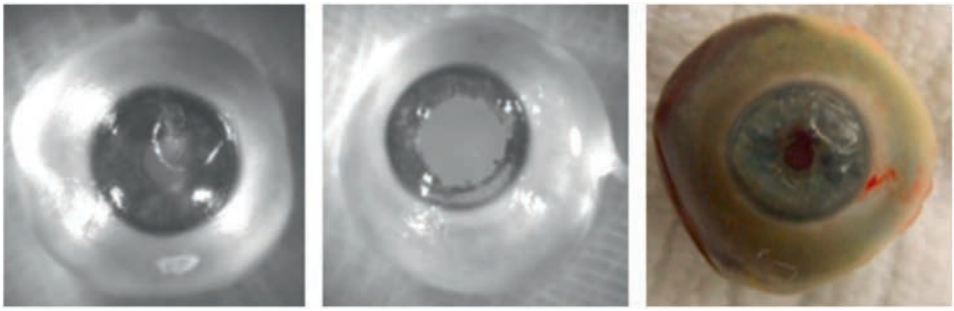


Fig. 1: (left to right): Deflation due to IOP, torn iris due to trauma, Tache Noire (in visible spectrum).

post-mortem human iris tissue include difficulties with obtaining and managing donated eyes for time-lapse postmortem studies.

With recent discoveries in postmortem iris biometrics, the role of iris as a forensic tool has gained the attention of researchers. Trokielewicz, Czajka and Maciejewicz pioneered postmortem iris recognition in humans in 2016 [MAP16a, MAP16b]. They have since published numerous articles providing an excellent account on postmortem iris biometrics.

Maciejewicz et. al. [MAP20] describe a method which employs a learning-based segmentation algorithm followed by a matching algorithm based on Gabor filters. They report equal error rates (EER) of less than 1% compared to 16.89% from off the shelf methods such as OSIRIS. These results are based on postmortem samples that are captured up to ten hours after death.

Maciejewicz et. al. [MAP19] analyzed the biometric capability of iris under different ocular pathologies. They showed the effects of different eye pathologies such as cataract on iris segmentation. They also show the biometric EER of irises with cataract are slightly greater than those of the normal irises.

In [MAP18], Maciejewicz et. al. discuss the effects of aging, disease and postmortem changes on the human iris. They show that differences in pupil dilation, combined with certain quality factors of the sample image and the progression of age as well as postmortem duration can significantly degrade the recognition accuracy.

Bolme et. al. in [Da16] discuss the effects of environment, including temperature, on natural decay of deceased bodies. The research collected a multi-modal postmortem biometric dataset and analyzed the feasibility of using fingerprint, face and iris under long term postmortem scenario. It was observed that postmortem fingerprints were easier to acquire and had lower false reject ratse compared to other postmortem biometric modalities.

Sansola et. al. [Al15] used IriShield M2120U iris recognition camera together with off the shelf IriCore matching software in their postmortem experiments involving 43 deceased subjects who had their irises photographed at different post-mortem time intervals. The study reported 19-30% false rejects and no false accepts. The study also reported on the

relationship between eye color and iris match scores, with blue-gray eyes yielding lower correct match rates than brown eyes.

4 Qualitative Analysis of Postmortem Iris

Researchers have studied qualitative changes related to ante to post mortem human ocular tissues including iris, retina, aqueous humor, and sclera [Da55]. Postmortem ocular changes, although not visible to the unaided eye, start early and on the onset of death. These processes are initially at molecular and cellular levels, and slowly progress into macroscopic level. Some of the affected tissues and the instigating factors related to postmortem iris recognition performance are described below.

Iris: The eye, when regarded as an extension of the body proper, can also be noted to exhibit the effects of trauma and systemic medical disease at the time of death. Several examples of trauma include injuries that could have led to the death of the subject, as well as post-surgical changes common in the elderly population (glaucoma surgery utilizing iridotomy or cataract surgery). These factors could impact the use of iris for biometrics. Complicating matters is the possibility of interval change in iris positioning and density caused by varying degrees of physiologic pupillary dilation.

Cornea: One of the most noticeable postmortem finding is corneal opacity. The cornea, although an extremely thin tissue, is particularly susceptible to changes in the environment. Corneal clarity is achieved by tightly organized collagen fibers with a specific percentage of hydration. If corneal hydration is altered by processes like surface epithelial drying or dysfunctional endothelial pump cells which could occur following death, corneal opacity can develop (Figure 2). In [JN94], the researchers observed that corneal opacity occurs approximately after two hours of death. Interestingly, the study also indicates that corneal opacity depends on season of death, i.e. postmortem opacity is increasingly seen in summer compared to winter season which is further corroborated by [Da16].

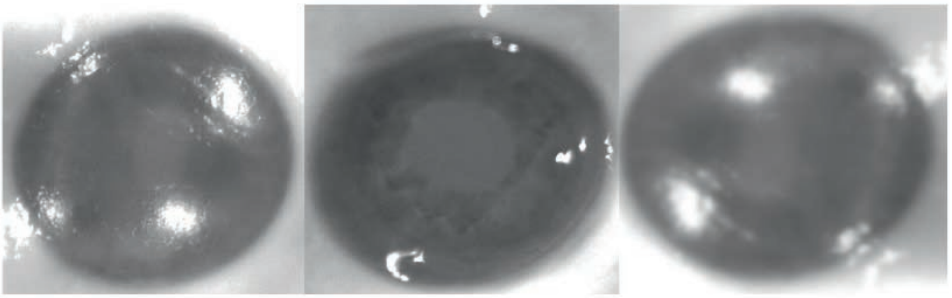


Fig. 2: (left to right): Pupil fadeout, limbic boundary diffusion, and corneal opacity.

Other Aspects *Tache Noire:* Tache Noire de la sclerotique is a phenomenon where conjunctivae darkens due to drying when the eyes remain open postmortem, creating darker

patches on the white of the eye [Pr03]. This process may make the eyeball rigid and and initiate deformation of the eyeball.

Limbic boundary fadeout: Limbic boundaries are circular and sharp. In the later stages of post-mortem analysis, these boundaries fade out due to opacification and natural degradation, making it difficult even for a human observer to determine the accurate limbic boundaries (Figure 2).

Intraocular Pressure (IOP): Loss of ocular tension in postmortem eyes causes the eye to flatten and lose its roundness, much like a deflated ball. This alters the shape and contour of the anterior segment of the eye, including the iris, and can negatively affect biometric utility [GDA15]. Change of intraocular pressure in postmortem studies is poorly understood due to a paucity of investigation.

5 Quantitative Study

Methods To analyze the postmortem utility of off the shelf iris recognition, we used a method similar to Open Source for IRIS (OSIRIS), which is an academic software developed within the BioSecure EU project [Su12]. It follows the original work of John Daugman, for iris image segmentation and subsequent normalization to a dimensionless polar coordinate system. A binary iris code is calculated using phase quantization of the Gabor filters. Hamming distance is used to compute dissimilarity score between two iris templates. Hamming distance between two irises 'a' and 'b', whose iris codes are codeA (Ca) and codeB (Cb), respectively, and whose valid segmentation masks are A and B, respectively, is given by:

$$d = \frac{\|(C_a \otimes C_b) \cap A \cap B\|}{\|A \cap B\|} \tag{1}$$

A match threshold of 0.42 was used in our study based earlier-mentioned control dataset. Since the goal of our experiment was to validate the effectiveness of iris recognition algorithms on postmortem enucleated iris, we manually corrected all limbic and pupil boundaries that were erroneous. We also removed iridial glare.

Experiment Setup We captured images in both visible and infrared spectra, and divided our quantitative experiments into three analyses:

1. Short Term Analysis (Unaided): For this analysis, data was captured during a six to seven hour observation period in one hour intervals. The enucleated eyes were left to naturally degrade at around 17 degrees Celsius. The aforesaid duration for short time analysis was experimentally determined based pilot observations on the first specimen, noting when it degraded to a point where iris tissue was completely deformed. This time is referred to as minimum decay time.
2. Long Term Analysis: After the minimum decay time, we continued imaging each specimen on a two-hour interval basis for two days, with two days being the longest possible duration after which the last specimen was entirely deformed.

3. Long Term Analysis with Hydration: While working on long term analysis, we observed that adding saline drops on specimen and storing the eyeballs in saline liquid helped the iris to retain its biometric value compared to when it was left out at room temperature. To further validate this observation, we stored a few specimens in saline under controlled temperature. These specimens were taken out only for iris captures. The iris was captured on an hourly basis for six hours per day for three days. In this setting, when the eyeball started to deform, we injected saline into the eyeball to retain its shape. This method was successful only until a certain iteration and stage, after which the eyeball could not retain its shape. This was especially evident on the onset of Tache Noire.

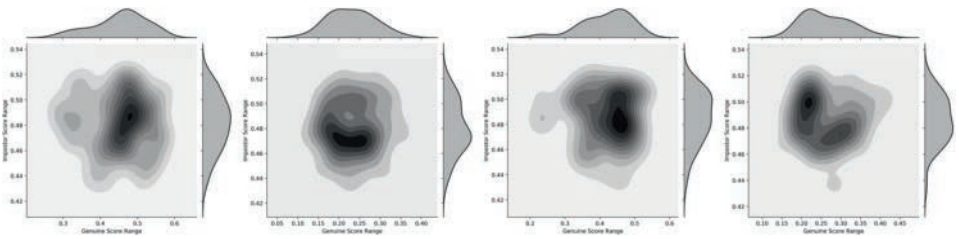


Fig. 3: (left to right) genuine (x-axis) and impostor (y-axis) score distributions for unaided long-term, unaided short term, aided long-term, and aided short-term analyses.

Quantitative Performance Analysis Figure 3 shows the distribution of genuine and impostor scores for different experiment settings. We report False Reject Rate (FRR) at 0% False Accept Rate (FAR) operating point, which was derived from the control subject data, as a measure of iris performance for the forthcoming analysis. The algorithm did reach the 0% FAR point for all the three postmortem experiment settings at such threshold, albeit with at varying FRRs.

1. Short term analysis: Short term analyses for both aided and unaided long-term study were similar. The enrollment template was computed from the first capture, and the remaining captures were used as verification samples. For the unaided study, the average FRR at 0% FAR was 3.25%. Similarly, for the aided study, the average FRR at 0% FAR was 0%. Figure 3 shows the distribution of match scores for this analysis. It was observed that the minimum and maximum FRRs for short term analysis were 0% and 12.5%, respectively. The specimen that caused higher FRRs had developed opacity at a faster rate than the other specimen. Also, it was noted that the iris colors of the specimen with larger FRR tended to be lighter.

2. Long term analysis (Unaided): Figure 5 shows the performance in terms of iris match scores with respect to postmortem duration for unaided long term analysis. The enrollment template was from corresponding short-term analysis experiment. Captures after the end of short term study were used as verification samples. The average FRR at 0% FAR was 78.47% for this experiment. Figure 3 shows the distribution of match scores for unaided short and long-term analysis. The minimum and

maximum FRRs for for this dataset was observed to be 62.5% and 91.6%, respectively. Six of the eleven specimen were used for this study.

3. Long term analysis with Hydration (Aided): Figure 4 shows the performance of long term analysis of eyes kept in saline solution (aided). The enrollment template was from corresponding short-term analysis experiment. Captures after the end of short term study were used as verification samples. The average FRR at 0% FAR was 55.8% for this long-term analysis. Figure 3 shows the distribution of match scores for this analysis. The minimum and maximum FRR for this dataset were 25.2% and 78%, respectively. Five of the eleven specimen were used for this study.

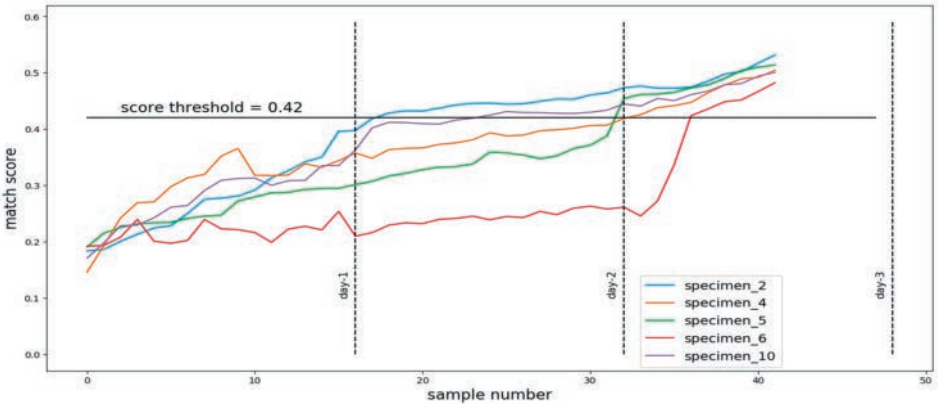


Fig. 4: iris match score progression with respect to aided short and long term postmortem period. .

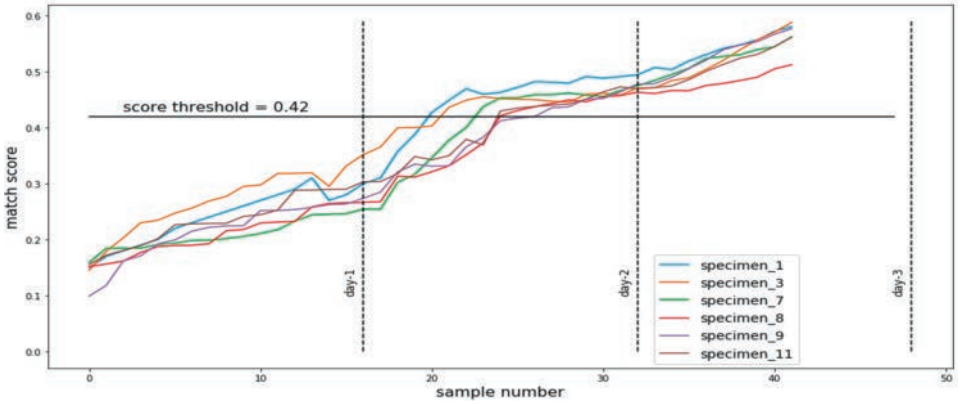


Fig. 5: iris match score progression with respect to unaided short and long term postmortem period.

6 Conclusion

This study answered many of our questions vexed at the start of our experiments, albeit with certain limitations.

First, we showed that off the shelf commercial iris recognition software can be used to identify enucleated cadaver eyes using normal (live population) thresholds. It was observed that the FRR of postmortem specimen can vary anywhere between 0% to 71% based on the postmortem period of the enucleated eyes, while FAR remains at 0. However, as shown by [SM11], the accuracy of traditional iris recognition methods can be further improved by using newer learning-based methods. Per the scope of this study, we did not evaluate postmortem iris recognition with such learning-based template extraction and matching algorithms.

We analyzed the effect of postmortem time lapse on enucleated eyes in terms of match score degradation under varying conditions. We observed the effect of environment and eye pressure (IOP) under different experimental setting on iris recognition performance. It was observed that retaining the shape of the eyeball by way of saline injections helped improving the genuine accept rates of the iris specimen, prolonging the biometric utility of postmortem iris tissue.

In [MAP18], the authors mention that a variant of OSIRIS with manual segmentation can reliably identify a postmortem iris in NIR spectrum up to 263 hours. In their study, iris was captured in-vivo; i.e. the eyeball was not enucleated. In our case, we could match enucleated iris up to 164.5 hours. However, in our study the time lapse is an average of (a) time of death to enucleation (b) enucleation of eye to date of first capture (c) time of first capture to loss of biometric value. It is interesting to note that only a few eyeballs started to deform under controlled storage mechanism. In our case, only two of sixteen rejected irises were due to deformed eyeball. However, while at room temperature, enucleated eyeballs deteriorated at varying rates, and in some cases they lost their iris textures in as little as eight hours.

Acknowledgement

This work was supported in part by a grant from Center for Identification Technology Reseach (CITeR), an NSF IUCRC. The authors wish to thank Dr. Arun Ross for his contributions and valuable comments on the early draft of this paper. We also would like to thank Dr. Gerald Early and Dr. Rohit Krishna for their valuable insights while designing this study. We would like to thank Heartland Lions Eye Bank for providing the enucleated eyeballs. Lastly we wish to thank Duc Huy Hoang (Mark) Nguyen for his assistance in editing and formatting of this manuscript. Dr. Derakhshani is also a consultant for Eyeverify (dba ZOLOZ).

References

[Al15] Alora, S.K.H.: Postmortem iris recognition and its application in human identification. PhD thesis, Master's Thesis, Boston University, 2015.

[Da55] Davson, H.: The hydration of the cornea. Biochemical Journal, 59(1):24, 1955.

[Da16] David, B.S.; Ryan, T.A.; Chris, B.C.; Tiffany, S.B.; Kelly, S.A.; Wolfe, S. Dawnie: Impact of environmental factors on biometric matching during human decomposition. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–8, 2016.

[EMD08] Elizabeth, A.; Margaret, C.; David, Q.: Pig-mentation: Postmortem Iris Color Change in the Eyes of Sus scrofa. Journal of forensic sciences, 53(3):626–631, 2008.

[GDA15] Gemma, P.; D, P. Maria; Aurelio, L.: Morphological and histological changes in eye lens: Possible application for estimating postmortem interval. Legal Medicine, 17(6):437–442, 2015.

[JD01] John, D.; D.Cathryn: Epigenetic randomness, complexity and singularity of human iris patterns. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1477):1737–1740, 2001.

[JN94] Jaafar, S.; Nokes, L.D.M.: Examination of the eye as a means to determine the early postmortem period: a review of the literature. Forensic science international, 64(2-3):185–189, 1994.

[MAP16a] Mateusz, T.; Adam, C.; Piotr, M.: Human iris recognition in post-mortem subjects: Study and database. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–6, 2016.

[MAP16b] Mateusz, T.; Adam, C.; Piotr, M.: Post-mortem human iris recognition. In: 2016 International Conference on Biometrics (ICB). IEEE, pp. 1–6, 2016.

[MAP18] Mateusz, T.; Adam, C.; Piotr, M.: Iris recognition under biologically troublesome conditions-effects of aging, diseases and post-mortem changes. arXiv preprint arXiv:1809.00182, 2018.

[MAP19] Mateusz, T.; Adam, C.; Piotr, M.: Iris recognition in cases of eye pathology. In: Biometrics under Biomedical Considerations, pp. 41–69. Springer, 2019.

[MAP20] Mateusz, T.; Adam, C.; Piotr, M.: Post-mortem iris recognition with deep-learning-based image segmentation. Image and Vision Computing, 94:103866, 2020.

[Pr03] Prasad, B.K.: Post-mortem ocular changes: a study on autopsy cases in Bharatpur Hospital. Kathmandu University medical journal (KUMJ), 1(4):276–277, 2003.

[SM11] Stephen, M.L.; Michael, M.L.: Document Title: Rapid Visualization of Biological Fluids at Crime Scenes using Optical Spectroscopy. 2011.

[Su12] Sutraand, G.; Dorizzi, B.; Garcia-Salicetti, S.; Othman, N.: A biometric reference system for iris. OSIRIS version 4.1. Telecom Sud Paris, France, Tech. Rep, 2012.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 139

Explaining ECG Biometrics: Is It All In The QRS?

João Ribeiro Pinto[1], Jaime S. Cardoso[1]

Abstract: The literature seems to indicate that the QRS complex is the most important component of the electrocardiogram (ECG) for biometrics. To verify this claim, we use interpretability tools to explain how a convolutional neural network uses ECG signals to identify people, using on-the-person (PTB) and off-the-person (UofTDB) signals. While the QRS complex appears indeed to be a key feature on ECG biometrics, especially with cleaner signals, results indicate that, for larger populations in off-the-person settings, the QRS shares relevance with other heartbeat components, which it is essential to locate. These insights indicate that avoiding excessive focus on the QRS complex, using decision explanations during training, could be useful for model regularisation.

Keywords: Biometrics, Electrocardiogram, Explainability, Identification, Interpretability.

1 Introduction

Throughout the past twenty years, research on biometrics based in the electrocardiogram (ECG) has largely been a success story [PCL18]. After successful proofs-of-concept in cleaner medical signals (*on-the-person*), the focus is quickly shifting to acquisitions in more realistic scenarios (*off-the-person*). Deep learning approaches [La18, Lu18, PCL19, PC19, Ha20] have been essential in dealing with the increased noise and variability in off-the-person settings, despite the performance and robustness issues that still hinder application in real scenarios.

However, deep learning decisions are obscure: unlike traditional methods based on fiducial features, we don't know what information the model uses to distinguish people. One can assume that the models look mainly to the QRS, since it is the most stable part of the ECG in the face of noise and variability [Sc00, HUvO01]. Several methods have thus focused on QRS complexes for ECG biometrics [Wa16, La18], but this practice has become uncommon in recent works. This indicates the true role of this waveform complex in identity discrimination is still to be adequately recognised.

Currently, pattern recognition researchers understand the importance of knowing what specific information is relevant for their models to reach decisions. Retreating to easily explainable traditional models (such as decision trees) is often unacceptable due to their performance limitations. Hence, various interpretability tools are being developed to peek into the inner workings of deep networks applied to diverse tasks [CPC19, SFC19, Se20].

This work uses, for the first time in the literature, such interpretability tools on a deep ECG biometric model, to understand what parts of the ECG are most useful for automatic human

[1] INESC TEC and Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, joao.t.pinto@inesctec.pt

identification. The model is a competitive state-of-the-art method [PCL19, PC19] applied for ECG-based identification in data subsets with diverse signal quality and number of identities. With this, we aim to assert the importance of the QRS and other waveforms for ECG biometrics and discuss future possibilities as this topic evolves towards more challenging and realistic scenarios. Additionally, we propose an intuitive way to visualise interpretations for unidimensional signals. The code and additional results are available online[2].

Besides this introduction, this paper presents some fundamental concepts on the ECG as a biometric trait, in section 2. The biometric identification model, the interpretability tools, and the visualisation method are described in section 3, and the experimental settings are detailed in section 4. Section 5 presents the obtained results and their discussion, and section 6 states the conclusions drawn from this work.

2 The Electrocardiogram as a Biometric Trait

The heart is composed of a muscle, the myocardium, that is responsible for its contraction and allows it to fulfil its purpose of pumping blood throughout the body [Ta09]. The myocardium contracts in response to depolarisation phenomena started by the atrioventricular node located on the interatrial septum. The waves of depolarisation that spread precisely across the heart are small electrical currents that can be measured using electrodes, resulting in the electrocardiogram (ECG) [MH13, Ta09].

Since the operation of the heart is a repetition of a sequence of phenomena, the ECG is approximately a cyclical repetition of a set of waveforms (P, Q, R, S, and T) that corresponds to a heartbeat (see Fig. 1) [MH13, PCL18]. The P wave is the first waveform and corresponds to the depolarisation of the myocardium cells in the atria. The Q, R, and S waveforms are commonly jointly considered as the QRS complex, which corresponds to the repolarisation of the atria and the depolarisation of the ventricles. The T wave corresponds to the repolarisation of the ventricles. This last wave is in some cases followed by a shorter waveform, the U wave, whose causes are still unclear [Ri08].

As a measurement of the electrical currents spread across the heart, the ECG signals will reflect the geometry of this organ. For example, larger hearts, with more cells to depolarise and repolarise, will result in ECG waveforms with larger amplitudes. Higher or lower basal heart rates will also result in different signal morphologies. Since heart geometry and basal heart rates vary across individuals, this intersubject variability is what makes the ECG sufficiently unique to be used in biometric recognition [HUvO01, vOHU00].

However, the ECG signals are also susceptible to intrasubject variability factors. Noise sources during acquisition, the short-term and long-term effects of exercise, emotional states, stress, drowsiness, and fatigue are some of the factors that reflect mainly in the heart rate variability, changing the morphology of the P-R and S-T segments [Sc00, ABH12]. These are the sources of uncertainty that hinder the use of the ECG as a biometric trait.

[2] xECG Repository. Available on: `https://github.com/jtrpinto/xECG`.

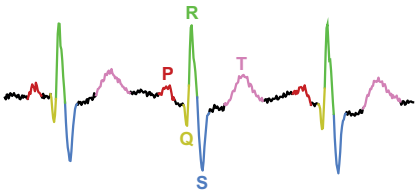


Fig. 1: Illustration of the ECG waveforms on a sample PTB signal segment.

While these are largely controlled on medical or on-the-person settings (where the subject is at rest, laying down, and signals are acquired using several high-quality gel electrodes), their effects are dominant for realistic off-the-person signals (acquired using few dry electrodes on the hands, during common daily activities) [Pi17, PCL18, PC19].

When compared with the P and T waves, the QRS corresponds to a larger polarisation event over a shorter period. In practice, this makes the QRS more dominant over noise and intrasubject variability than the other ECG waveforms [Pi17, PCL18]. Hence, the QRS is considered more stable over time and across variable conditions, which makes it better suited for biometric recognition.

Despite this, it is still unclear how much identity information is carried by the QRS complex compared to the other waveforms, and whether it is enough for an accurate and robust biometric recognition system. Studies on ECG-based biometric identification have shown it is possible to distinguish small sets of individuals in on-the-person settings using only the QRS complex or QRS fiducial amplitude and time measurements [Wa16, La18]. Nevertheless, this practice is becoming uncommon as research evolves towards realistic off-the-person signals and larger databases.

This denotes that the sole use of the QRS may not be adequate for off-the-person settings, or the individual information carried by the QRS may not be enough to distinguish individuals in large populations. This work aimed to address these doubts through a study on the role and relevance of the QRS and the other waveforms on ECG-based biometric identification. Interpretability tools are used to assess which parts of the ECG are more relevant to the decisions of an end-to-end identification model [PCL19], with on-the-person and off-the-person signals and data subsets with a varying number of identities.

3 Methodology

3.1 Biometric Identification Model

The biometric model for identification followed the architecture proposed by Pinto *et al.* [PCL19], which has attained state-of-the-art results in off-the-person settings for both identification and, later, authentication [PC19]. The model (see Fig. 2) receives five-second blindly segmented ECG signals and outputs probabilities for each of the N identities considered. Finding the highest probability score allows us to assign the respective identity to the input signal.

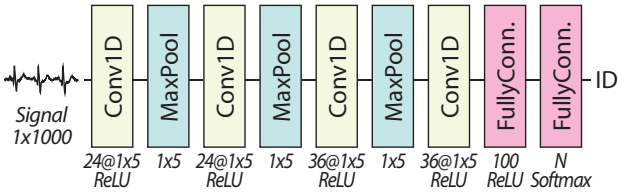


Fig. 2: Architecture of the biometric identification model

The model consists of an end-to-end 1D convolutional neural network (CNN) with four convolutional layers (with 1×5 filters, two layers with 24 followed by two with 36), followed by ReLU activation. Neighbouring convolutional layers are separated by 1×5 max-pooling layers. The last convolutional layer is followed by two fully-connected layers (100 neurons with ReLU and N neurons with softmax activation).

3.2 Interpretability Tools

To capture the dynamics behind the decisions of the biometric model, four interpretability methods are applied to the trained model: Occlusion, Saliency, Gradient SHAP, and DeepLIFT. Occlusion and Saliency are two of the simplest interpretability methods, while Gradient SHAP and DeepLIFT are more sophisticated and powerful. These are implemented in the Captum toolbox [Ko19] for PyTorch and are described below.

Occlusion The Occlusion method [ZF14] consists in measuring the influence of hiding a portion of the input on the output of the model. When hidden, the more relevant input parts will cause larger changes in the output, and will thus be assigned greater relevance in the explanations offered by this method. This is the simplest method to interpret a model, although the size of occluded regions should be carefully defined to obtain meaningful explanations.

Saliency The Saliency method [SVZ14] is based on the gradients of a model given a certain input. Through backpropagation, the gradient of target class scores w.r.t. the input is obtained. A saliency map is then generated by rearranging the class score derivatives, generating saliency maps that assign higher relevance to input regions that correspond to higher gradients. Requiring a single backpropagation pass, this method is a simple and fast way to obtain explanations on model predictions.

Gradient SHAP Gradient SHAP [LL17] is an approach based on game theory which considers the explanations of a model's predictions as models themselves. For sophisticated deep learning models, the explanation models are simplified and interpretable approximations of the respective models. SHapley Addictive exPlanation (SHAP) values, inspired by game theory's Shapley values, are computed through the gradient of a random

point between a baseline and the input with added random noise. The SHAP values denote how much a given part of the input raises the probability for the considered class, and are reportedly better aligned with human intuition and effective in discriminating among output classes.

DeepLIFT DeepLIFT (Deep Learning Important FeaTures) [SGK17] performs back-propagation to track the contributions to the output to the responsible parts of the input. Throughout this process, it compares the difference in inputs and outputs considering a reference (or baseline) input, assigning contribution scores to each neuron of the model. It also allows for the study of negative contributions: how much a specific part of the input contributes to lower the probability for the considered class.

3.3 Visualisation

Decision explanations obtained using interpretability tools are visualised using the multi-coloured line plot feature of Matplotlib [Hu07]. ECG signals are plotted so that the colour of each signal component represents its relative relevance for the decision. In this case, lighter yellow colours represent less relevant time samples, whereas more relevant samples assume darker purple colours. This way, both the ECG morphology and the relevance of each of its components are easily and intuitively presented.

4 Experimental Settings

The data used for model training and evaluation have been drawn from the Physikalisch-Technische Bundesanstalt ECG Database (PTB) [BKS95, Go00] and the University of Toronto ECG Database (UofTDB) [Wa14]. The PTB database includes on-the-person (high-quality) 12-lead ECG signals acquired at 1 kHz from 290 subjects at rest. The UofTDB includes single-lead off-the-person (more noisy and realistic) data acquired from 1019 subjects. To match the UofTDB, PTB signals were downsampled to 200 Hz and only Lead I was used.

Five-second segments were blindly extracted (without fiducial detection) from the recordings. Fifty per cent of those segments (*per* identity) were used during training and the remaining were reserved for testing. This provided more challenging test settings than those commonly found in the literature, but also deliberately avoided the most realistic settings (see [PC19]), for the sake of obtaining meaningful interpretations.

To simulate gradually increasing identification difficulty within each database, subsets of N identities are considered, with $N \in \{2, 5, 10, 20, 50, 100, 200, 500, 1019\}$. The identities in each subset are the first N in lexicographical order. Each subset includes all identities that compose smaller subsets, so subjects #1 and #2 are the main focus of analysis since these are present in all subsets. Throughout this paper, T_N denotes the subset of UofTDB data from N subjects and P_N denotes the subset of PTB data from N identities. As stated in

Tab. 1: True positive identification rate results (%) on the test data.

Database	Number of Identities								
	2	5	10	20	50	100	200[1]	500	1019
PTB	100.0	100.0	99.63	99.50	98.92	98.76	97.73	-	-
UofTDB	100.0	97.26	98.30	95.46	93.86	91.16	89.70	91.20	91.45

[1]For PTB, this column corresponds to the entire set of 290 subjects.

Table 1, P_{290} was used instead of P_{200} to take advantage of the entire PTB dataset. Model training details can be found online at this project's repository.

Performance evaluation is based on the True Positive Identification Rate (or accuracy): the fraction of test samples that are correctly assigned to their true identity by the trained model. Interpretations are examined through the proposed visualisation method.

5 Results and Discussion

The results of the performance evaluation are presented in Table 1. These results roughly follow the expected patterns considering the use of on-the-person *versus* off-the-person ECG data. The model is able to attain high true positive identification rates in both databases when the population is small, but as the set of subjects grows, performance decreases and a wide gap distinguishes the more challenging off-the-person settings from the more controlled on-the-person settings.

Additionally, one can find some unusual patterns in the performance results. Considering $M > N$, one would expect identification performance with subset T_N to be higher than with subset T_M. With UofTDB off-the-person data this is not always verified: *e.g.*, from T_5 to T_{10}, performance increases from 97.26% to 98.10%. In these cases, we need to consider that datasets with fewer identities have fewer data and, thus, more unstable results. Alternatively, the identities added to T_N to create T_M may be easier to discriminate ("sheep", according to the concept of biometric menagerie [Do98, YD10]) and thus contribute to improve accuracy. However, one should also regard the substantial regularisation needed to avoid overfitting and the instability during training as possible causes for these discrepancies. This is a very important insight into the increased difficulties of using off-the-person data and the need for improved and more robust biometric models.

Analysing the explanations obtained using the four interpretability tools (examples in Fig. 3 and Fig. 4), a trend is verified from smaller to larger identity subsets, consisting on the deviation from focusing mainly on the QRS complex to the increasing relevance of other parts of the heartbeats. This is also confirmed when combining the explanations of all heartbeats of each person into a single average heartbeat (see Fig. 5 and Fig. 6).

With the cleaner medical signals from PTB, the focus is mostly on the QRS complex, but information from other waveforms starts to become more and more relevant as more identities are added. It is noteworthy how, when discriminating PTB subjects #1 and #2 in a two-subject scenario (see Fig. 5), the model still focuses mainly on the QRS, even

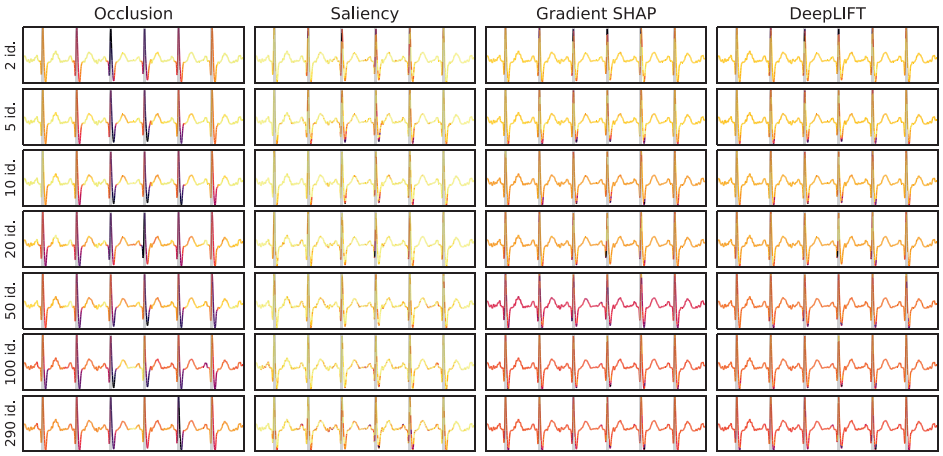


Fig. 3: Explanations over an example five-second ECG segment from PTB. In each subplot, the yellow to dark purple colours correspond to increasing time sample relevance and vertical grey lines denote R-peak locations. Signals were filtered for easy visualisation.

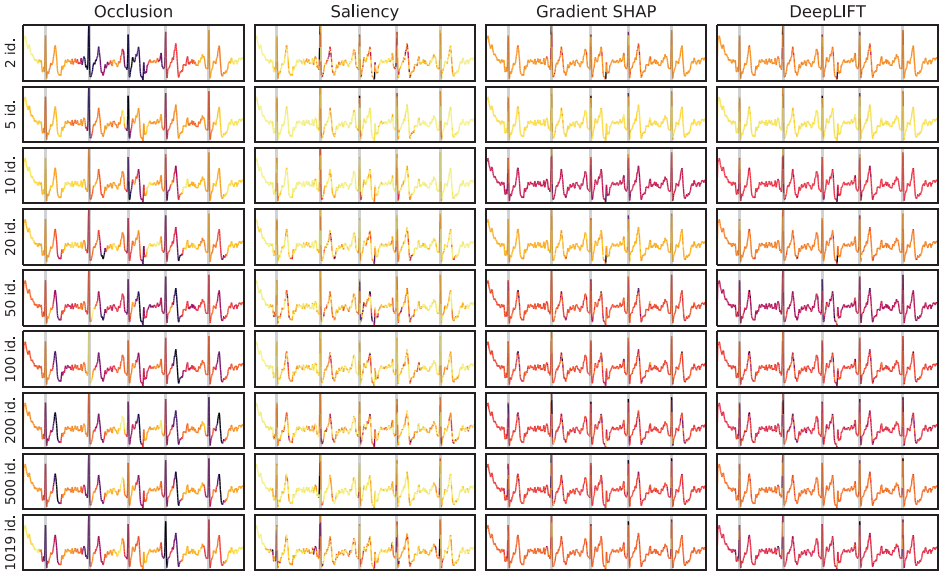


Fig. 4: Explanations over an example five-second ECG segment from UofTDB. In each subplot, the yellow to dark purple colours correspond to increasing time sample relevance and vertical grey lines denote R-peak locations. Signals were filtered for easy visualisation.

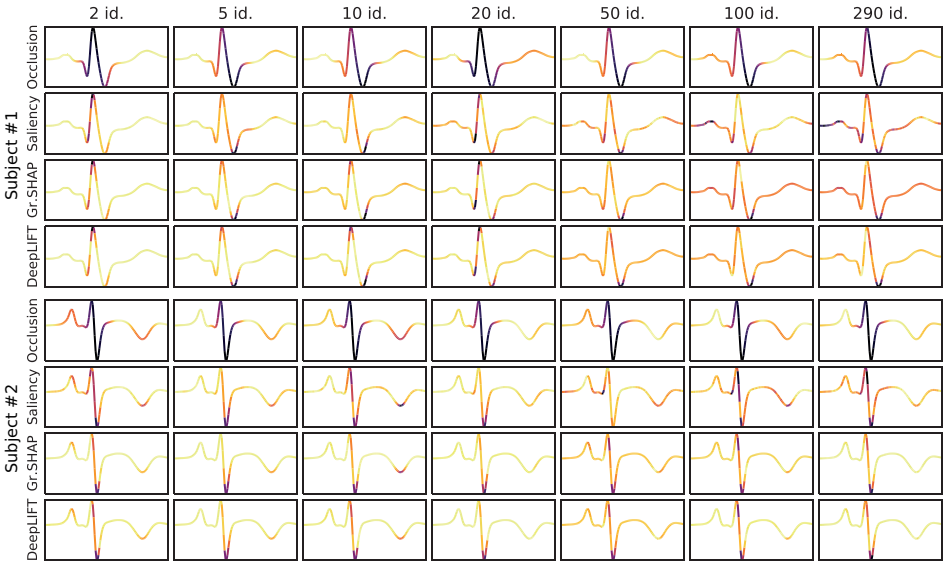


Fig. 5: Average explanations over heartbeat waveforms of subjects #1 (top) and #2 (bottom) on the subsets of the PTB database. In each subplot, the yellow to dark purple colours correspond to increasing time sample relevance. Signals were filtered for easy visualisation.

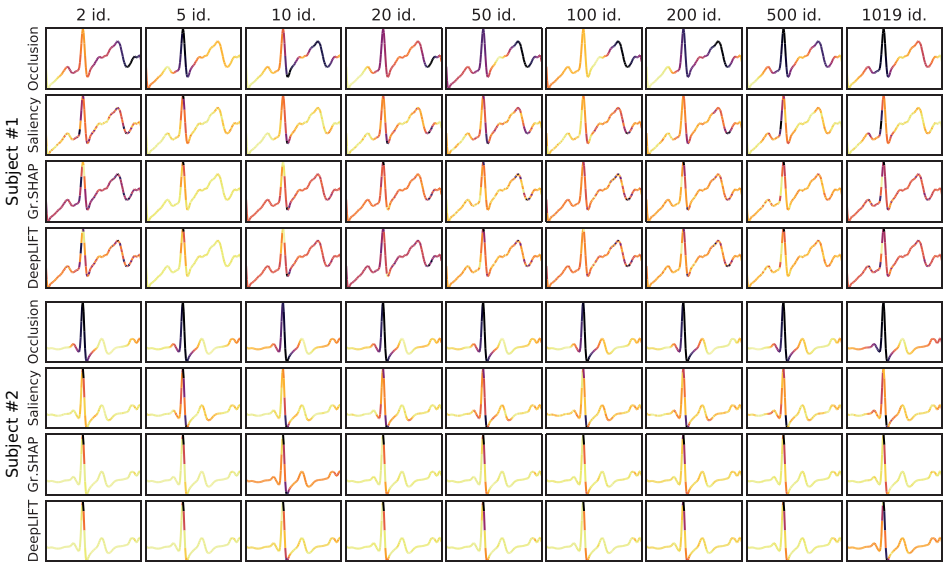


Fig. 6: Average explanations over heartbeat waveforms of subjects #1 (top) and #2 (bottom) on the subsets of the UofTDB database. In each subplot, the yellow to dark purple colours correspond to increasing time sample relevance. Signals were filtered for easy visualisation.

though subject #2 has a very specific characteristic, the inverted T-wave, that is arguably their most distinctive feature. This denotes how, in these cleaner signals, the QRS complex is so stable that the remaining waveforms, more susceptible to heart rate variability, are largely ignored by the model regardless of any visually obvious intersubject differences they may present.

With the more realistic off-the-person signals from UofTDB, the QRS retains high importance but the relevance is more evenly spread among the signal waveforms. In the specific case of subject #2 (see Fig. 6), it is evident that the QRS retains the highest importance for the decision, even in T_{1019} (the largest subset). This may denote that, even in these more challenging settings, the identification models will still give preference to the QRS over other waveforms if it is sufficiently unique among the considered identities. Nevertheless, in such large sets of identities, the expected behaviour is that of subject #1 (see Fig. 6), since the limited identity information carried by the QRS will lead the model to also look to other parts of the signal.

One interesting aspect is the difference between the results with Occlusion *versus* the other methods. Occlusion generally grants the QRS complex much more relevance, regardless of the settings. In the state-of-the-art approaches, the QRS complex is not only a source for identity features but also frequently used as an easily detectable reference landmark for the location of other ECG waveforms. This may also be the case in this end-to-end deep model. Although there are challenging contexts where the QRS may not be the main contributor to the decision, it may be essential to the deep model as a reference landmark to locate other waveforms in the signal. Hence, when occluded, it will be the signal component that most impacts the decision, causing the occlusion method to generally consider it the most relevant.

6 Conclusion

This work aimed to explain how deep models use ECG signals to distinguish people, using interpretability tools. Overall, the obtained results partially confirm the claim that the QRS is the key to ECG-based biometrics. With small populations in on-the-person settings, it can alone be used for reliable recognition. However, as we evolve towards larger populations and off-the-person settings, other components become relevant in discriminating people, as the models require more identity information to overcome the hurdles placed by enhanced intrasubject variability.

However, even though relevance is more evenly shared in off-the-person identification in large sets of identities, the QRS is shown as essential by the occlusion method. It appears that, just like several literature methods, the implemented end-to-end model learnt to use the QRS as a landmark for the location of other ECG components in the signal, resulting in large output changes when the QRS is occluded. Hence, despite the literature claims, one should avoid relying too heavily on any single part of the ECG, including the QRS complex, since all waveforms carry identity information that proves increasingly useful in more realistic settings and larger populations.

Beyond these insights, further efforts should be devoted to extend this study and offer a deeper, more thorough, and more objective analysis of the contribution of each ECG waveform to the model's decisions. Obtaining more systematic and complete explanations could create new opportunities on the use of interpretability tools during model training. Using explanations to regularise models and promote focus in the most relevant signal components or the distributed use of the whole signal (instead of just the QRS) could lead to improved recognition accuracy and robustness.

Acknowledgements

This work was financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalization - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project "POCI-01-0145-FEDER-030707", and within the PhD grant "SFRH/BD/137720/2018". The authors wish to thank the creators and administrators of the PTB (Physikalisch-Technische Bundesanstalt, Germany) and UofTDB (University of Toronto, Canada) databases, which have been essential for this work.

References

[ABH12] Agrafioti, F.; Bui, F. M.; Hatzinakos, D.: Secure Telemedicine: Biometrics for Remote and Continuous Patient Verification. Journal of Computer Networks and Communications, 2012:11, 2012.

[BKS95] Bousseljot, R.; Kreiseler, D.; Schnabel, A.: Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik, 40(1), 1995.

[CPC19] Carvalho, D. V.; Pereira, E. M.; Cardoso, J. S.: Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, 2019.

[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, National Institute of Standards and Technology, 1998.

[Go00] Goldberger, A.; Amaral, L.; Glass, L.; Hausdorff, J.; Ivanov, P. C.; Mark, R.; Stanley, H. E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23):e215–e220, 2000.

[Ha20] Hammad, M.; Pławiak, P.; Wang, K.; Acharya, U. R.: ResNet-Attention model for human authentication using ECG signals. Expert Systems, p. e12547, 2020.

[Hu07] Hunter, J. D.: Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3):90–95, 2007.

[HUvO01] Hoekema, R.; Uijen, G. J. H.; van Oosterom, A.: Geometrical aspects of the interindividual variability of multilead ECG recordings. IEEE Transactions on Biomedical Engineering, 48(5):551–559, May 2001.

[Ko19] Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Reynolds, J.; Melnikov, A.; Lunova, N.; Reblitz-Richardson, O.: , PyTorch Captum. https://github.com/pytorch/captum, 2019.

[La18] Labati, R. D.; Muñoz, E.; Piuri, V.; Sassi, R.; Scotti, F.: Deep-ECG: Convolutional Neural Networks for ECG biometric recognition. Pattern Recognition Letters, 126:78–85, 2018.

[LL17] Lundberg, S. M.; Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems. pp. 4765–4774, 2017.

[Lu18] Luz, E. J. S.; Moreira, G. J. P.; Oliveira, L. S.; Schwartz, W. R.; Menotti, D.: Learning Deep Off-the-Person Heart Biometrics Representations. IEEE Transactions on Information Forensics and Security, 13(5):1258–1270, May 2018.

[MH13] Marieb, E. N.; Hoehn, K.: Human Anatomy & Physiology. Pearson, Glenview, IL, ninth edition, 2013.

[PC19] Pinto, J. R.; Cardoso, J. S.: An End-to-End Convolutional Neural Network for ECG-Based Biometric Authentication. In: 10th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). Tampa, FL, United States, 2019. (in press).

[PCL18] Pinto, J. R.; Cardoso, J. S.; Lourenço, A.: Evolution, Current Challenges, and Future Possibilities in ECG Biometrics. IEEE Access, 6:34746–34776, June 2018.

[PCL19] Pinto, J. R.; Cardoso, J. S.; Lourenço, A.: Deep Neural Networks For Biometric Identification Based On Non-Intrusive ECG Acquisitions. In: The Biometric Computing: Recognition and Registration, chapter 11, pp. 217–234. CRC Press, 2019.

[Pi17] Pinto, J. R.; Cardoso, J. S.; Lourenço, A.; Carreiras, C.: Towards a Continuous Biometric System Based on ECG Signals Acquired on the Steering Wheel. Sensors, 17(10), 2017.

[Ri08] Riera, A. R. P.; Ferreira, C.; Filho, C. F.; Ferreira, M.; Meneghini, A.; Uchida, A. H.; Schapachnik, E.; Dubner, S.; Zhang, L.: The enigmatic sixth wave of the electrocardiogram: the U wave. Cardiology Journal, 15(5):408–421, 2008.

[Sc00] Schijvenaars, R. J. A.: Intra-individual Variability of the Electrocardiogram: Assessment and exploitation in computerized ECG analysis. PhD thesis, Erasmus University Rotterdam, 2000.

[Se20] Sequeira, A. F.; Silva, W.; Pinto, J. R.; Goncalves, T.; Cardoso, J. S.: Interpretable Biometrics: Should We Rethink How Presentation Attack Detection is Evaluated? In: 2020 8th International Workshop on Biometrics and Forensics (IWBF). 2020.

[SFC19] Silva, W.; Fernandes, K.; Cardoso, J. S.: How to produce complementary explanations using an Ensemble Model. In: 2019 International Joint Conference on Neural Networks (IJCNN). 2019.

[SGK17] Shrikumar, A.; Greenside, P.; Kundaje, A.: Learning Important Features through Propagating Activation Differences. In: 34th International Conference on Machine Learning. pp. 3145–3153, 2017.

[SVZ14] Simonyan, K.; Vedaldi, A.; Zisserman, A.: Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: Workshop at International Conference on Learning Representations. 2014.

[Ta09] Tate, P.: Seeley's Principles of Anatomy and Physiology. McGraw-Hill, New York, NY, second edition, 2009.

[vOHU00] van Oosterom, A.; Hoekema, R.; Uijen, G. J. H.: Geometrical factors affecting the interindividual variability of the ECG and the VCG. Journal of Electrocardiology, 33:219 – 227, 2000.

[Wa14] Wahabi, S.; Pouryayevali, S.; Hari, S.; Hatzinakos, D.: On Evaluating ECG Biometric Systems: Session-Dependence and Body Posture. IEEE Transactions on Information Forensics and Security, 9(11):2002–2013, Nov 2014.

[Wa16] Waili, T.; Nor, R. M.; Rahman, A.; Sidek, K. A.; Ibrahim, A. A.: Electrocardiogram Identification: Use a Simple Set of Features in QRS Complex to Identify Individuals. In: 12th International Conference on Computing and Information Technology (IC2IT). pp. 139–148, 2016.

[YD10] Yager, N.; Dunstone, T.: The Biometric Menagerie. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(2):220–230, February 2010.

[ZF14] Zeiler, M. D.; Fergus, R.: Visualizing and Understanding Convolutional Networks. In: 2014 European Conference on Computer Vision (ECCV). pp. 818–833, 2014.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 151

Action-Independent Generalized Behavioral Identity Descriptors for Look-alike Recognition in Videos

Ali Khodabakhsh[1], Hugo Loiselle[2]

Abstract: There is a long history of exploitation of the visual similarity of look-alikes for fraud and deception. The visual similarity along with the application of physical and digital cosmetics greatly challenges the recognition ability of average humans. Face recognition systems are not an exception in this regard and are vulnerable to such similarities. In contrast to physiological face recognition, behavioral face recognition is often overlooked due to the outstanding success of the former. However, the behavior of a person can provide an additional source of discriminative information with regards to the identity of individuals when physiological attributes are not reliable. In this study, we propose a novel biometric recognition system based only on facial behavior for the differentiation of look-alikes in unconstrained recording conditions. To this end, we organized a dataset of 85,656 utterances from 1000 look-alike pairs based on videos collected from the wild, large enough for the development of deep learning solutions. Our selection criteria assert that for these collected videos, both state-of-the-art biometric systems and human judgment fail in recognition. Furthermore, to utilize the advantage of large-scale data, we introduce a novel action-independent biometric recognition system that was trained using triplet-loss to create generalized behavioral identity embeddings. We achieve look-alike recognition equal-error-rate of 7.93% with sole reliance on the behavior descriptors extracted from facial landmark movements. The proposed method can have applications in face recognition as well as presentation attack detection and Deepfake detection.

Keywords: Behavioral Biometrics, Face Recognition, Look-alike face, Facial Motion, Triplet Loss.

Fig. 1: Examples of look-alike identity pairs in the proposed 1000 look-alike pairs (1000LP) dataset. Each column shows one pair of look-alikes. The identities in the proposed dataset are a subset of the identities in the VGGFace2 [Ca18] dataset.

1 Introduction

Distinguishing visually similar individuals, be it identical twins or look-alikes with physical make-up or plastic surgery, has been challenging for both humans and face recognition

[1] NTNU, IIK, Norwegian Biometrics Lab, Gjovik, NO, ali.khodabakhsh@ntnu.no
[2] Orange, Région de Caen, Fr, loiselle.hugo@hotmail.fr

algorithms [La11]. In the context of video communication, this vulnerability is further exacerbated as other means of identity verification are often not available. Moreover, the use of look-alikes and make-up for fraud has an advantage over digital manipulation methods as they don't produce any digital footprint in the received signal to be used for detection. Furthermore, despite the rise of advanced digital video manipulation methods such as Deepfakes, subjective tests show higher susceptibility of viewers to fake videos containing look-alikes rather than digitally manipulated videos [KRB19]. Fortunately, a video signal contains additional clues on the identity of the person in the form of facial behavior [Be10, KJ97].

Among existing methods for behavioral face recognition (BFR), the vast majority of studies focus on fixed-phrase authentication or specific emotional responses. Chen et al. [LLJ01] propose use of dense optical flow vector distance for identification in a fixed-phrase scenario. In [Ce06] Cetingul et al. experiment with dense motion features, lip contour motion features, and lip shape features with a hidden-Markov-model (HMM) classifier. Zafeiriou and Pantic [ZP11] use principal component analysis (PCA) followed by linear discriminant analysis (LDA) on dense facial deformation features in spontaneous smile for biometric recognition. Wang and Liew [WL12] show that behavioral lip biometrics based on temporal shape descriptors and motion vector representation outperforms physiological lip biometrics based on texture descriptors. Gavrilescu [Ga16] proposes a multi-state neural network on individual facial expressions extracted in the form of facial action coding system (FACS). More recently, Iengo et al. [Ie19] use neural networks on dynamic facial features to achieve a fixed-phrase recognition rate of 98.2% and Taskirar et al. [Ta19] use statistical properties of facial distances during different phases of smile facial expression for face recognition.

A number of publications have attempted to address unconstrained BFR. Matta and Dugelay [MD06] propose using rigid head displacements along with GMM and Bayesian classifiers for person recognition. Ye and Sim [YS10] use locally similar facial deformation patterns for identification through the calculation of local deformation profile similarity. In [Sh16], Shreve et al. quantify the type and intensity as well as the temporal dynamics of action units (AU) via calculating histogram distances and dynamic time warping (DTW) distance. Yuan et al. [Yu17] propose the usage of active shape models on lip contour along with gaussian mixture models (GMM) for authentication in smartphone applications.

BFR has also been used in multi-modal biometric recognition as well as presentation attack detection (PAD). Notably, Zhao and Pietikanien [ZP07] introduce local binary patterns (LBP) on three orthogonal planes and volume LBPs and thus incorporates immediate neighborhood frames of the video for face recognition. Kim et al. [KKR16] use long short-term memory (LSTM) cells on top of convolutional neural networks (CNN) to capture smile facial dynamics. More recently, Pan and Deravi [PD17] use support vector machine (SVM) on AU histogram features for presentation attack detection. Finally, Agrawal et al. [Ag19] model facial expressions of four individuals using facial landmarks and SVM to detect Deepfakes.

To distinguish look-alikes from each other many image-based methods have been proposed. Klare et al. [KPJ11] provide a taxonomy of facial features and analyze the dis-

criminative power of these features for identical twin identification. The only video-based solution is proposed by Zhang et al. [Zh14], where they extracted six types of face motion from the talking profile of identical twins and use the similarity of aligned motion sequences for classification by an SVM model. To the best of the authors' knowledge, there exists no publicly available video dataset of look-alikes in the literature. The only related video dataset in the literature is the private dataset by Zhang et al. [Zh14] collected from 39 pairs of twins at the Mojiang International Twins Festival. There also exists a couple of related datasets containing solely images. Lamba et al. [La11] collected the only dataset on look-alikes consisting of 500 images from 50 celebrities and their look-alikes. Phillips et al. [Ph11] collected a dataset of 435 twins consisting of 24050 images.

All aforementioned publications rely on small data collected in controlled environments, and few of them address emotion- and utterance-independent detection with limited success, and as such, among all publications regarding this topic, none have addressed the unconstrained BFR in real-world scenarios. In this study, we introduce a general-purpose action-independent identity descriptor extractor based on facial behavior for distinguishing look-alikes. To this end, we also provide the first large-scale look-alike video dataset named "1000 look-alike pairs (1000LP)" which consists of approximately 23,000 real-world videos collected from a public video-sharing platform[3], for which both humans and state-of-the-art recognition systems fail at differentiation[4]. Among the aforementioned literature, the approach in this article is in the same line of research as is taken by Zhang et al. [Zh14] and Agrawal et al. [Ag19]. The rest of this article is organized as follows: in Section 2 the proposed method is described, while Section 3 includes the details of the collected dataset as well as the experiment setup. The results of the experiments are discussed in Section 4 and the article is concluded in Section 5.

2 Proposed Method

The physiological likeliness of two individuals due to natural similarity or application of physical or digital makeup may lead to false-positives in face recognition. In these cases, the facial behavior can be a source of complementary information for face recognition. Facial behavior contains identifiable information and has a significant role in person identification by humans [Be10, KJ97]. In our proposed method, after face detection and facial landmark extraction in each frame, we train a convolutional deep neural network (CDNN) which maps the sequence of normalized landmark positions in the video to a vector in a generalized behavior space in an end-to-end manner. This approach enables the recognition of persons that are previously unseen by the detector by simply calculating the distance between behavior-vectors extracted from a pair of videos. Furthermore, as the network only sees the landmarks, it is guaranteed to be void of influence by the physiological likeliness of the individuals. Furthermore, landmarks are not as sensitive to disturbances and quality-related issues as other features such as optical and motion fields are and can be extracted with higher confidence.

[3] http://www.youtube.com

[4] The dataset is publicly available for download at `http://ali.khodabakhsh.org/research/1000lp/`

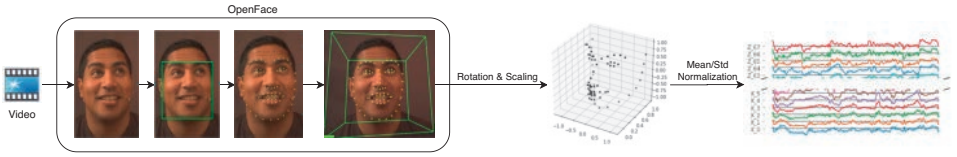


Fig. 2: Feature extraction pipeline.

2.1 Preprocessing

We use the open-source facial behavior analysis toolkit OpenFace [BRM16] to extract the landmark positions from videos. The toolkit provides face detection as well as pose estimation and 3D landmark positions for each frame in the video. For landmark positions to be independent of the camera position and head rotation angle, we use the pose estimation information to rotate the 3D landmarks in 3D space to achieve a frontal pose of zero degrees roll, yaw, and pitch. Further on, the landmark positions in each video are scaled to match a fixed scale used over the whole dataset. The scaling is done such that the inner eye corner landmarks would be on average 0.5 units apart. Finally, the landmarks are individually normalized using their mean and standard deviation across the whole training dataset. The aim of the aforementioned normalization steps is to convert the landmark positions to rotation-independent displacements from the average position. Even though the pose information can also contain additional behavioral identity information, they were left out due to their dependence of the estimated pose to the camera angle and position. Figure 2 visualizes the preprocessing pipeline.

2.2 The proposed recognition system

To extract identity-sensitive yet action-independent information from the time series of landmark movements, it is fruitful to rely on the distribution statistics of the landmark deviations. However, due to the noisy nature of the estimated 3D landmark positions extracted from 2D videos in the pre-processing step, a refinement step proves necessary. However, the refinement criteria are ambiguous as the correct landmark position is not available. Furthermore, the movements are correlated to a large extent and contain redundancies. Motivated by the recent success of x-vectors [Sn18] in the field of speaker recognition, we propose the network architecture shown in Figure 3 for end-to-end learning of the appropriate refinement for the best identification performance before statistical pooling. In this architecture, four 1D-convolutional layers are applied to the input time series. By using max-pooling layers across time, the receptive field of the final layer of the stack can be increased. Following the convolutional layers, a linear mapping is learned to map the output of the last convolutional layer to the feature-embedding space. After calculation of the mean and standard deviation of the feature-embeddings across time, the resulting fixed-length vector is then used for generating identity embeddings by two fully-connected layers. Instead of using class labels for training the network, we use triplet loss [SKP15] to enable better generalization capacity for unseen identities. Furthermore, batch normalization is used after the input layer, the statistical pooling layer, and between the output of

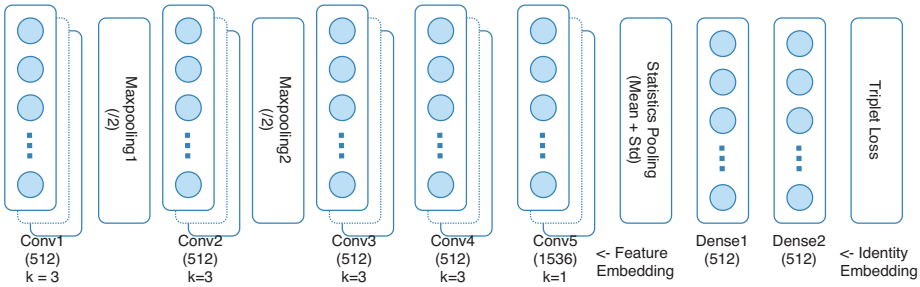


Fig. 3: Proposed network architecture.

neurons and activation functions to reduce the learning time of the network. No activation is used on the output of the feature-embedding mapping layer and the final layer to enable the network to utilize the full embedding space.

The Euclidean distance between identity embeddings can directly be used as a biometric dissimilarity metric. In the case of multiple enrollment samples from multiple identities, it is also possible to use the proposed system as a preprocessing step, and train a softmax layer for classification directly on extracted identity embeddings.

2.3 Look-alike mining

The VoxCeleb2 dataset [CNZ18] contains over 1 million utterances from more than 6,000 celebrities collected from YouTube. The identities in this dataset are a subset of identities in the VGGFace2 [Ca18] dataset. To mine for Look-alike identities, we used the ArcFace [De19] face recognition system to compare the average embeddings for each identity in the VGGFace2 dataset that appears in VoxCeleb2 dataset as well. After sorting the scores of the resulting 36M comparison pairs, the top 2,000 pairs with the highest similarity score are selected for a subjective face recognition test. Among the top pairs, there exist pairs of identical twins as well.

In the subjective face recognition test, for each look-alike pair of identities, four images are selected from each identity from the VGGFace2 dataset and shown to participants. The task for the participants was to check whether the two sets of images correspond to the same identity or two different people. The user interface is shown in Figure 4. Due to the large number of comparisons, the test was done by 20 participants, each labeling 200 pairs such that each pair is labeled by two people. From the resulting comparisons, the pairs that were labeled as the same people by at least one participant were selected as look-alikes and formed the 1000 look-alike pairs (1000LP) dataset. Figure 1 shows examples of the resultant look-alike pairs. To assure the reliability of the selected look-alike pairs, the equal-error-rate (EER) is calculated for the resulting look-alike pairs using the ArcFace network, resulting in an unacceptably high EER of 30.32%.

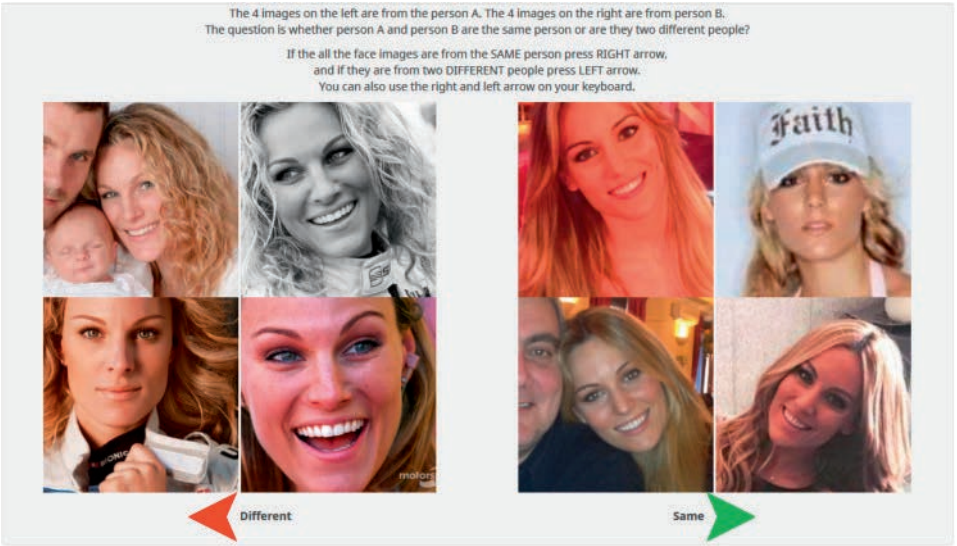


Fig. 4: Subjective face recognition test user interface.

3 Experiment Setup

The selected 1000 pairs of look-alikes consist of 1634 unique identities. The remaining 4500 identities in VoxCeleb2 are available for training the network. The rest of this section describes the details of the organized test dataset and the parameters used for training.

3.1 1000LP Dataset

The utterances available in the VoxCeleb2 dataset are in the format of cropped faces sized 224×224 pixels at 25 frames per second in AVC1 format. There is a total of $1,128,246$ utterances which originate from $150,480$ YouTube videos. After filtering out all utterances with a length of less than 8 seconds and discarding all utterances for which face landmark detection failed, a total of $253,361$ utterances remained for training and $85,656$ utterances for testing. The median length of the remaining utterances is 10.7 seconds. From the 4500 train identities, 15% of them were held for validation purposes, and the remaining were used for training. For the test identities, one-third of videos ($28,368$ utterances) were separated for enrollment, and the remaining videos were used for testing. Resulting from this, $127,332$ test trials were created, out of which $57,288$ are client trials and $70,044$ are impostor trials[5]. Special care is taken in the selection of the enrollment and test utterances such that if an utterance from a YouTube video is used in the enrollment, no utterances from the same video remains in the test trials. Thus, the performance is assured to correspond to the cross-video performance in real-life use.

[5] The dataset is publicly available for download at `http://ali.khodabakhsh.org/research/1000lp/`

		Verification	Identification	
		EER (%)	Top-1 (%)	Top-5 (%)
Euclidean Distance	Segment (∼10 sec)	15.42	60.57	77.83
	Video (∼4 seg)	13.04	**79.84**	**92.61**
Softmax Classifier	Segment (∼10 sec)	10.08	65.47	81.00
	Video (∼4 seg)	**7.93**	73.87	86.33

Tab. 1: The performance of the proposed methods.

3.2 Detector

The network parameters are shown in Figure 3. The breadth of the network along with the dimension of the final embedding is set to 512, with only the exception of expanded feature embedding dimensions of three times the breadth. The total number of trainable parameters in the network was 5.3M. A kernel size of 3 is used in the convolutional stack while max-pooling is done with a stride of two, resulting in a receptive field of 23 frames (roughly one second) before statistical pooling. The normalized input had a dimension of 204 corresponding to 3D coordinates for the 68 landmarks. The model was trained using TensorFlow[6] with a batch size of 256 and the learning rate was manually adjusted towards minimizing validation loss. Semi-hard triplet loss on L2 distance of L2 normalized network outputs was used and the model was trained for 10 epochs. The hyper-parameters are selected according to the highest network performance on validation data.

4 Results and Discussion

The verification and identification performance of the proposed method for Euclidean similarity as well as softmax probabilities are reported in Table 1. The Euclidean similarity scoring performs better in identification mode than softmax probabilities and achieves an identification accuracy of 79.84% on video level. This is remarkable considering the large number of identities enrolled in the system (1634). Despite the high identification accuracy, the EER of the Euclidean similarity measure is 13.04%. Softmax probabilities, however, achieve a much better EER of 7.93% in verification mode. This discrepancy shows that softmax probabilities perform better in separating score distributions of client and impostor trials, but fails to preserve the ranking order of similarities. The detection error tradeoff (DET) curve is shown in Figure 5 visualizing the fact.

In order to be able to interpret the performance of the proposed method, it is compared to the reported results for existing BFR methods in the literature in Table 2. It is important to emphasize that all previous methods have only been tested on videos with controlled and semi-controlled recording environments. Among the methods that operate on non-predetermined motion, the proposed method has the lowest EER and a comparable recognition rate despite the number of enrolled identities being orders of magnitude larger.

[6] https://www.tensorflow.org/

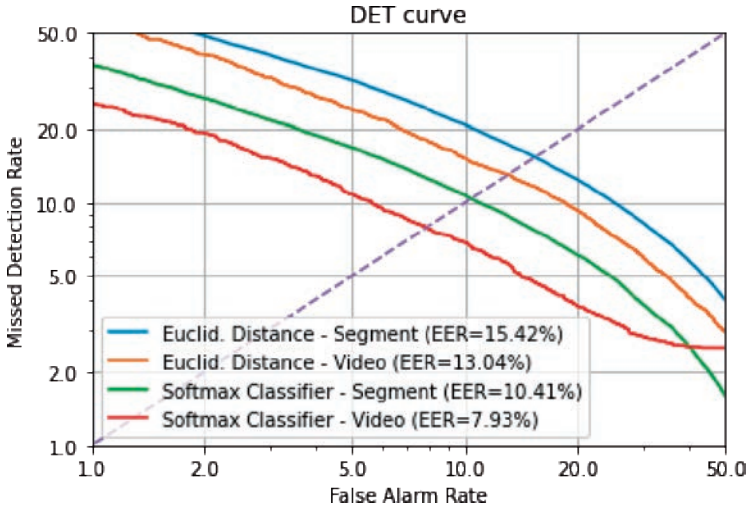


Fig. 5: Detection error tradeoff (DET) curve for the proposed methods.

Ref.	Subj. #	Environment	Motion	Feature	Classifier	Perf.	Metric
[LLJ01]	28	Controlled	Fixed-Phrase Speech	Motion Flow Fields	PCA + LDA	∼87%	Recog. Rate
[Ce06]	50	Controlled	Fixed-Phrase Speech	Grid-based Motion Contour-based Motion Lip Shape	LDA + Bayes	5.2% 12.0% 10.4%	EER
[ZP11]	22	Controlled	Spontaneous Smile	Motion Fields	PCA + LDA	2.5%	EER
[WL12]	40	Controlled	Fixed-Phrase Speech	Lip Shape Deformation Lip Texture Deformation	HMM-UBM	1.92% 8.53%	EER
[Ga16]	64	Controlled	Induced Emotion	Facial Action Units	MLP	91.7%	Percision
[Ie19]	20	Controlled	Fixed-Phrase Speech	Facial Landmarks	DNN	0.64%	EER
[Ta19]	400	Controlled	Spontaneous Smile	Facial Landmark Distances	Euclid. Dist.	31.20%	EER
[MD06]	9	Controlled	Unconstrained Speech	Facial Feature Displacement	GMM	19.1%	EER
[YS10]	97	Controlled	Unconstrained Emotion	Local Deformation Patterns	Similarity	18.86%	EER
[Sh16]	96	Ambiguous	Unconstrained Speech	Facial Action Units	Hist. Sim. DTW	∼62%	Recog. Rate
[Yu17]	20	Ambiguous	Unconstrained Speech	Lip Contour	GMM	96.2%	Recog. Rate
[Ag19]	Clinton Sanders Trump Warren	Ambiguous	Unconstrained Speech	Facial Action Units	SVM	75% 95% 77% 95%	TPR @ 10% FPR
Proposed	1634	Unconstrained	Unconstrained Speech	Facial Landmarks	CDNN	7.93% 79.84% 93.12%	EER Recog. Rate TPR@ 10% FPR

Tab. 2: The performance of the proposed method in contrast to the reported results for existing methods.

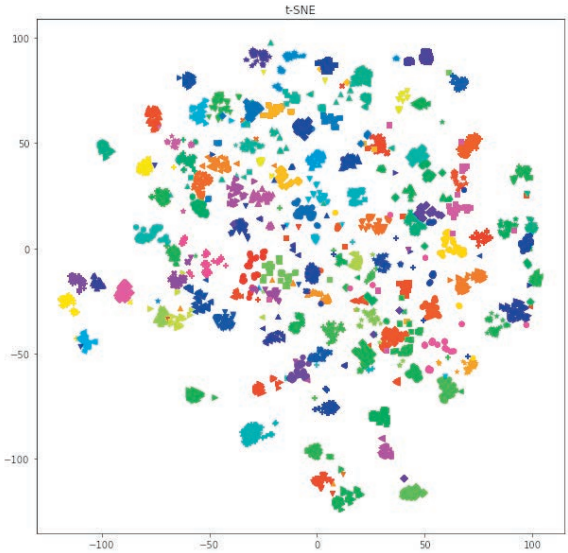


Fig. 6: t-distributed stochastic neighbor embedding for enrollment utterances. For aesthetic reasons, only the identities with more than 50 enrollment utterances are visualized. Different colors and shapes signify different identities.

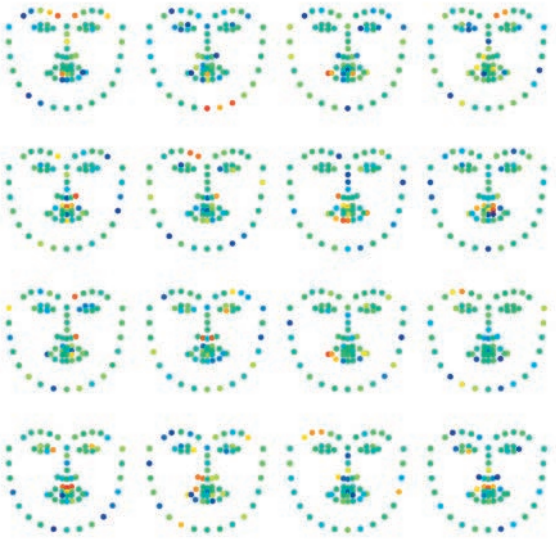


Fig. 7: Facial landmark significance visualization for selected filters in conv1. The significance is measured as the norm of the 3×3 matrix corresponding to x, y, and z coordinates of the landmark in frames $t-1$, t, and $t+1$.

The t-distributed stochastic neighbor embeddings (t-SNE) [MH08] for enrollment utterances for a subset of identities is visualized in Figure 6. It is visible that the enrollment utterances of test set identities form concentrated clusters with few outliers. This signifies that the learned embedding space is able to generalize well across unseen identities, and the failure cases probably correspond to the outliers. Figure 7 shows landmark significance for a selected set of filters in the first convolutional layer of the network. The significance is measured in terms of the norm of the 3×3 matrix corresponding to multiplicative weights in x, y, and z coordinates of each landmark in frames $t-1$, t, and $t+1$. These heatmaps show the reliance of the network on meaningful facial actions such as eyebrow movements, upper lip movements, and movements in the corners of the mouth.

The results of this study show the power of large data in improving the performance and generalizability of BFR systems. Even though this system is trained on 4500 identities, the number of training identities is still much smaller compared to physiological face recognition systems, and there is room for further improvement.

5 Conclusion

In this article, we proposed a novel general-purpose action-independent behavioral identity embedding extraction network with acceptable performance for real-life applications. The network benefits from a large number of training samples and identities and proves capable of extracting descriptive embeddings for unseen identities in unconstrained conditions. We also respond to the lack of publicly available large-scale datasets for look-alike detection, as well as publicly available behavioral face recognition systems by releasing the 1000 look-alike pairs (1000LP) dataset and the code for the proposed method.

The proposed method provides a complementary source of identity information that can be used alongside physiological face recognition systems to make them robust against look-alikes, as well as presentation attacks that try to mimic the physiological likeliness. The proposed method is robust to physical and digital spatial signal manipulations as it relies solely on the temporal behavior of the individual in question. Due to the permanence of behavioral face biometrics [Be10] and its robustness to manipulations and quality degradation, these methods have already found their way into the detection of Deepfakes [Ag19] and can provide a robust alternative to existing narrowly applicable detection methods [Kh18].

References

[Ag19] Agarwal, Shruti; Farid, Hany; Gu, Yuming; He, Mingming; Nagano, Koki; Li, Hao: Protecting World Leaders Against Deep Fakes. In: CVPR Workshops. June 2019.

[Be10] Benedikt, L.; Cosker, D.; Rosin, P. L.; Marshall, D.: Assessing the Uniqueness and Permanence of Facial Actions for Use in Biometric Applications. IEEE SMCS, 2010.

[BRM16] Baltrušaitis, T.; Robinson, P.; Morency, L.: OpenFace: An open source facial behavior analysis toolkit. In: WACV. pp. 1–10, 2016.

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

[Ce06] Cetingul, H. E.; Yemez, Y.; Erzin, E.; Tekalp, A. M.: Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading. IEEE TIPS, 2006.

[CNZ18] Chung, Joon Son; Nagrani, Arsha; Zisserman, Andrew: VoxCeleb2: Deep Speaker Recognition. CoRR, abs/1806.05622, 2018.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. pp. 4685–4694, 2019.

[Ga16] Gavrilescu, M.: Study on using individual differences in facial expressions for a face recognition system immune to spoofing attacks. IET Biometrics, 5(3):236–242, 2016.

[Ie19] Iengo, D.; Nappi, M.; Ricciardi, S.; Vanore, D.: Dynamic Facial Features for Inherently Safer Face Recognition. In: ICIP. pp. 2611–2615, 2019.

[Kh18] Khodabakhsh, A.; Ramachandra, R.; Raja, K.; Wasnik, P.; Busch, C.: Fake Face Detection Methods: Can They Be Generalized? In: BIOSIG. pp. 1–6, 2018.

[KJ97] Knight, Barbara; Johnston, Alan: The Role of Movement in Face Recognition. Visual Cognition, 4(3):265–273, 1997.

[KKR16] Kim, S. T.; Kim, D. H.; Ro, Y. M.: Facial dynamic modelling using long short-term memory network: Analysis and application to face authentication. In: BTAS. 2016.

[KPJ11] Klare, B.; Paulino, A. A.; Jain, A. K.: Analysis of facial features in identical twins. In: IJCB. pp. 1–8, 2011.

[KRB19] Khodabakhsh, A.; Ramachandra, R.; Busch, C.: Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content. In: QoMEX. pp. 1–6, 2019.

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

[LLJ01] Li-Fen Chen; Liao, H. . M.; Ja-Chen Lin: Person identification using facial motion. In: ICIP. volume 2, pp. 677–680 vol.2, 2001.

[MD06] Matta, F.; Dugelay, J.: A Behavioural Approach to Person Recognition. In: 2006 IEEE International Conference on Multimedia and Expo. pp. 1461–1464, 2006.

[MH08] Maaten, Laurens van der; Hinton, Geoffrey: Visualizing data using t-SNE. Journal of machine learning research, 9(Nov):2579–2605, 2008.

[PD17] Pan, S.; Deravi, F.: Facial action units for presentation attack detection. In: EST. pp. 62–67, 2017.

[Ph11] Phillips, P. J.; Flynn, P. J.; Bowyer, K. W.; Bruegge, R. W. V.; Grother, P. J.; Quinn, G. W.; Pruitt, M.: Distinguishing identical twins by face recognition. In: Face and Gesture 2011. pp. 185–192, 2011.

[Sh16] Shreve, M.; Bernal, E. A.; Li, Q.; Kumar, J.; Bala, R.: A study on the discriminability of facs from spontaneous facial expressions. In: ICIP. pp. 1674–1678, 2016.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: CVPR. June 2015.

[Sn18] Snyder, D.; Garcia-Romero, D.; Sell, G.; Povey, D.; Khudanpur, S.: X-Vectors: Robust DNN Embeddings for Speaker Recognition. In: ICASSP. pp. 5329–5333, 2018.

[Ta19] Taskirar, M.; Killioglu, M.; Kahraman, N.; Erdem, C. E.: Face Recognition Using Dynamic Features Extracted from Smile Videos. In: INISTA. pp. 1–6, 2019.

[WL12] Wang, Shi-Lin; Liew, Alan Wee-Chung: Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power. Pattern Recognition, 2012.

[YS10] Ye, N.; Sim, T.: Towards general motion-based face recognition. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2598–2605, 2010.

[Yu17] Yuan, Y.; Zhao, J.; Xi, W.; Qian, C.; Zhang, X.; Wang, Z.: SALM: Smartphone-Based Identity Authentication Using Lip Motion Characteristics. In: SMARTCOMP. 2017.

[Zh14] Zhang, Li; Ma, KengTeck; Nejati, Hossein; Foo, Lewis; Sim, Terence; Guo, Dong: A talking profile to distinguish identical twins. Image and Vision Computing, 2014.

[ZP07] Zhao, G.; Pietikainen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE TPAMI, 29(6):915–928, 2007.

[ZP11] Zafeiriou, S.; Pantic, M.: Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter. In: CVPR WORKSHOPS. pp. 13–19, 2011.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 163

3D Face Recognition For Cows

Deepak Yeleshetty[1], Luuk Spreeuwers[2], Yan Li[3]

Abstract: This paper presents a method to recognize cows using their 3D face point clouds. Face is chosen because of the rigid structure of the skull compared to other parts. The 3D face point clouds are acquired using a newly designed dual 3D camera setup. After registering the 3D faces to a specific pose, the cow's ID is determined by running Iterative Closest Point (ICP) method on the probe against all the point clouds in the gallery. The root mean square error (RMSE) between the ICP correspondences is used to identify the cows. The smaller the RMSE, the more likely that the cow is from the same class. In a closed set of 32 cows with 5 point clouds per cow in the gallery, the ICP recognition demonstrates an almost perfect identification rate of 99.53%.

Keywords: Cows, Biometrics, Visual identification, 3D face recognition, Pointcloud registration, Iterative Closest Point, Realsense cameras.

1 Introduction

Biometric identification is an efficient and a reliable method because it uses the unique natural discriminating features of each subject without the need of an external identification document or an attached device. This paper aims to design, implement, test and qualify a system that identifies cows using their 3D face point clouds. The project was carried out in the Product Development Group of the Dutch agri-tech company - Lely Industries N.V. Despite existing electrical cow identification methods, computer vision is opted due to its reliability, cost effectiveness and non-invasive property. Twisted Infrared (IR) tag around the neck sometimes results in failure of identification as the tag faces the cow's body. The IR sensor's batteries are non-replaceable, making the current system expensive. The tag around the neck also causes discomfort to the cow. Cameras are cost-effective and visual biometric identification methods like face recognition are non-invasive. The anatomy of the cow shows us that the skull is rigid and symmetric [JC07], which gives enough reason to pursue the cows' 3D faces for identification. This paper attempts to answer if we can uniquely identify cows based on their 3D face shapes. A new dual 3D camera setup is designed to capture the face of the cow. The 3D face is registered to a specific pose by finding the region of interest and correcting the rotation angles using the vertical symmetry of the cow's face [Sp11]. For recognition, the probe point cloud is compared with all the point clouds in the gallery using ICP [BM92] and inlier RMSE, a metric from the python library Open3D [ZPK18] is used to identify the cow. The identification rate for a herd of 32 cows is 99.53%, which proves that the 3D face shape can be used to identify cows.

[1] University of Twente, Data Management and Biometrics Group, Enschede, NL, yeleshetty.deepak@gmail.com
[2] University of Twente, Data Management and Biometrics Group, Enschede, NL, l.j.spreeuwers@utwente.nl
[3] Lely Technologies, Maassluis, NL, yli@lely.com

The paper is organized as follows: Related work on visual cow identification, human face recognition and point cloud registration methods is explained in section 2. The methodology of the proposed system is explained in section 3. The results are presented and analyzed in section 4. The paper is concluded with an insight on future scope in section 5.

2 Related Work

The authors of [Be19] demonstrated a Deep Learning method to identify cows based on multiple perspectives of their 2D face images. The accuracy for a closed set of 561 images from 52 cows was observed to be 89%. Their paper explains the shortcomings in terms of 2D landmark annotation, owing to the shape of the cow's face and states that multiple views yield better identification results.

ICP based recognition systems have been explored for 3D human faces [Ma05], however, as mentioned in [Sp11], ICP takes several seconds to register and recognize. The author in [Sp11] describes a fast and accurate 3D face registration and recognition method with a rank-1 identification rate of 99%. For 3D face registration, the region of interest (ROI) is estimated by fitting a cylinder. The vertical symmetry plane is obtained by finding the rotation around y and z axes. The angle between the nose bridge and the vertical axis is maintained at $\frac{\pi}{6}$ rad. Recognition is done by estimating the likelihood ratio of the probe's PCA-LDA features after comparing with those in the gallery. This method overcomes the time complexity of ICP and speeds up 3D face recognition for humans. As opposed to humans, cows lack the luxury of publicly available face database. Additionally, 2D face registration for cows is challenging as the face creates self occlusion for even a minute change of pose.

Two cameras are used in this project to overcome self occlusion. Point cloud registration is the process of estimating the rigid body transformation matrix that aligns the perspectives from both cameras, giving us a complete view of the subject. ICP [BM92] estimates the transformation between two point clouds (source to target) by minimizing the distances between correspondences, given an initial transformation. The transformation matrix is iteratively updated to minimize the point to point distances over the correspondence set. Let $C = \{(p,q)\}$ be the correspondence set with correspondence pairs $p \in P$ and $q \in Q$, where P and Q are the target and the source point clouds respectively. The two main ICP result metrics described in the library Open3D [ZPK18] are called Fitness (F) and Inlier Root Mean Square Error (I_{RMSE}).

$$F = \frac{N_c}{N_p} \qquad I_{RMSE} = \frac{1}{N_c} \sum_{(p,q)\in C} d_{p-q}$$

N_c is the number of correspondences, N_p is the number of points in the target point cloud and d_{p-q} is the mean squared distance between the correspondences. Fitness describes the overlapping area between the two point clouds. Inlier RMSE is the average of the mean square point to point distances of the correspondences(*Inliers*). A good registration results in a high fitness value (in the range [0,1]) and a low inlier RMSE value.

This paper aims to demonstrate the 3D face registration method explained in [Sp11], on cows. The recognition method will be based on point-to-point ICP [BM92].

3 Methodology

The proposed system's methodology can be divided into three steps: Data Acquisition, 3D Face Registration and ICP Based Recognition These steps can be seen in figure 1. The camera setup is designed to capture the 3D recordings of the cow's face. From the recordings, the required frames are captured and the point clouds are extracted. L-R registration method is performed on the extracted point cloud pairs to obtain 3D faces. The 3D faces are de-noised and transformed to a common pose as described in section 3.2 [Sp11]. Point to point ICP [BM92] is performed on every probe point cloud against all the point clouds in the gallery. The resulting inlier RMSE score is used to identify cows.

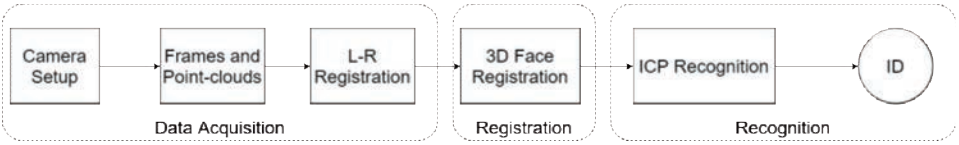


Fig. 1: Project Pipeline

3.1 Data Acquisition

A new dual 3D camera [4] setup is designed to acquire the 3D face of the cow. Both cameras are designed to face forward with no tilt because tilting the cameras would increase the field of view, which brings other cows in the frame and affects the further steps. Since the approximate ear to ear width of the cow was about 35 cm, the cameras were placed 70 cm apart. An illustration of the setup can be seen in figure 2, where C_L and C_R represent the left and right cameras respectively. F_W, F_D and S_B denote the approximate face width, approximate distance of the face from the setup and the fixed setup baseline respectively. Figure 3 shows the setup used in the farm.

Table 1 shows the specification of the camera setup and the camera itself. Due to the auto exposure setting in the Realsense camera, it was observed that the 3D points were very poorly estimated for cows with white fur or a surface that reflects light. So, only black or dark skinned cows are used in this project for identification. From a one-day data acquisition session, 1442 point clouds from 32 cows were collected and are used in this project. Each cow has 10 to 75 point clouds. Five point clouds per cow are stored in the gallery and the remaining are used as probe. The cows were treated gently without any discomfort throughout the data acquisition process.

To combine the point clouds from both cameras, L-R Registration method is followed (L-R indicates Left - Right cameras). L-R registration is divided into two steps: a feature-based

[4] Intel Realsense D435

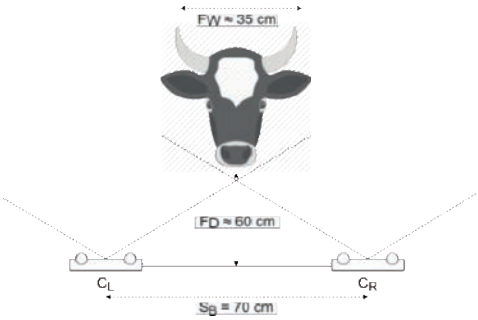


Fig. 2: Data Acquisition - Illustration

Fig. 3: Hardware set up

Description	Value
Approx. dist. cow to camera	60 cm
Setup baseline	70 cm
Cow face width	35 cm
Diagonal Field of view (per camera)	95°± 3°
Resolution	848 × 480 px
Frames per second	15 fps

Tab. 1: Camera Setup Specification

Description	Value
Voxel size for downsampling	0.02 m
Downsampling search radius	0.04 m
FPFH features search radius	0.1 m
RANSAC distance threshold	0.03 m

Tab. 2: Point Cloud Registration Parameters

Global Registration method that estimates a coarse transformation matrix and a *Local Registration* method that refines the transformation.

Fast Point Feature Histograms (FPFH) features are estimated on the down-sampled point cloud [RBB09]. A coarse transformation is obtained from the FPFH correspondences between the left (source) and the right (target) point clouds using RANdom SAmple Consensus (RANSAC) [FB81]. The RANSAC model is set to converge when the distance between majority of the correspondences reaches a global minimum. This coarse transformation matrix is fed to the ICP algorithm as an initial transformation estimate and yields a fine transformation matrix between the left and right cameras. This resulted in a visually convincing L-R Registration. Table 2 shows the different parameters used in L-R registration method. On an Intel i7 6-core 2.20 GHz CPU, it takes roughly 2 seconds to complete L-R Registration for one pair of point clouds.

3.2 3D Face Registration [Sp11]

3D face registration involves de-noising and transforming the L-R registered 3D face point cloud to a specific pose. The chosen pose is the front view of the cow, with the nose bridge

area parallel to the image plane, which results in an ideal perspective that shows the vertical symmetry of the cow's face. A slightly modified version of the face registration method explained in [Sp11] is implemented in this section.

To estimate the ROI of the point cloud, the surface normals of the point cloud are calculated and a cylinder is fit using RANSAC. The open source C++ library PCL[RC11] is used to fit a cylinder and extract the ROI using the defined parameters (Table 3) on the point cloud. The input and output of the ROI estimation is illustrated in figure 4. As opposed to

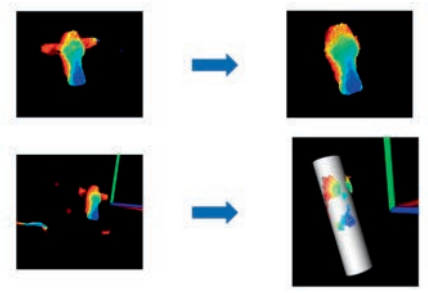


Fig. 4: Estimating ROI using PCL Cylinder Fitting

Description	Value
Radius Interval	$[0.15, 0.20]$ m
Inlier Distance Threshold (from Axis)	$radius \pm 0.05$ m
Max. RANSAC Iterations	1000

Tab. 3: Cylinder fitting parameters

humans, cows have a longer and relatively flatter nose bridge. So, a plane P_α with normal N_α is fitted on the ROI point cloud using RANSAC in PCL. This plane always fits on the nose bridge with a very minor tilt. The x-y plane is called P_{xy} with normal N_z. The angle γ between P_α and P_{xy} is calculated using their normals and the ROI is rotated around the x-axis by this angle. A plane $P_{n\alpha}$ is fit on the rotated ROI point cloud and it is translated along the positive z- axis to a distance $d_z = D - 0.1$ where D is the distance between the planes $P_{n\alpha}$ and P_{xy}. This will translate the point cloud approximately 10 cm from the x-y plane. To estimate the rotation angles along the y and z axes (θ and ϕ), we use the vertical

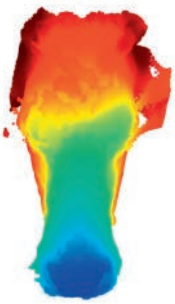


Fig. 5: Completely registered point cloud

Description	Value
θ interval	$[-\frac{\pi}{9}, \frac{\pi}{9}]$ rad.
θ step size	$\frac{\pi}{90}$ rad.
ϕ interval	$[-\frac{\pi}{4}, \frac{\pi}{4}]$ rad.
ϕ step size	$\frac{\pi}{50}$ rad.
Sliding interval	$[-\frac{3}{4}width, \frac{3}{4}width]$
Sliding step	0.05 m
$diff_z$ threshold	0.01

Tab. 4: Symmetric Orientation Parameters

symmetry of the cow's face and implement parts of the Symmetry Plane section in [Sp11]. A low resolution range image is constructed for multiple steps of rotation around the y axis (θ), by projecting the point cloud on the x-y plane with 5x5 mm grids. The value of each

pixel in this range image is equal to the average depth (z coordinate) of points projected to the corresponding grid. The image is rotated in-plane in multiple steps which is the same as rotating along z-axis (ϕ). For every step in ϕ, The range image is mirrored and is slid horizontally in $[-\frac{3}{4}w, \frac{3}{4}w]$ with a step size of $d = 5$ mm, where w is the width of the range image. For every step d, the pixel-wise difference ($diff_z$) between the image and its mirror is computed. The pixel is said to contribute to the symmetry if the $diff_z$ value lies below the threshold (0.01). The θ and ϕ step corresponding to the maximum number of contributing pixels are the required angles to *straighten* the cow's face. The result is called a completely registered point cloud (figure 5). The parameters used to obtain the symmetric orientations is summarized in the table 4. It was observed that some gallery point clouds are incorrectly registered but the source of these irregularities is not investigated in this project.

3.3 ICP Based Recognition

ICP based recognition method is identical to the two-step L-R registration method. It is a computationally expensive and time consuming process as each of the probe point cloud is compared with all 160 gallery point clouds (32 cows with 5 point clouds each). Figure 6 shows an overview of the ICP based recognition method. ICP on each probe generates 160 Fitness and Inlier RMSE scores. The inlier RMSE scores are grouped for each cow in gallery and the average scores per cow is computed, which results in a reduced set R_s of 32 scores. The gallery ID corresponding to the minimum inlier RMSE score of R_s is the predicted ID. Out of 1282 probes from 32 cows, 1276 probes are correctly predicted, yielding an identification rate of 99.532%. Recognizing each cow takes about 300 seconds on an Intel i7 6-core 2.20 GHz CPU. With further improvements in the data acquisition process and implementation of a version of [Sp11], the recognition process could be much faster.

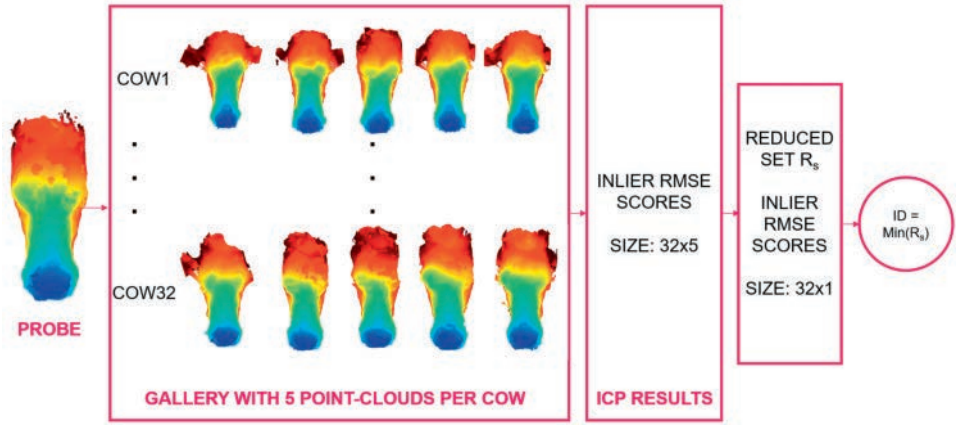
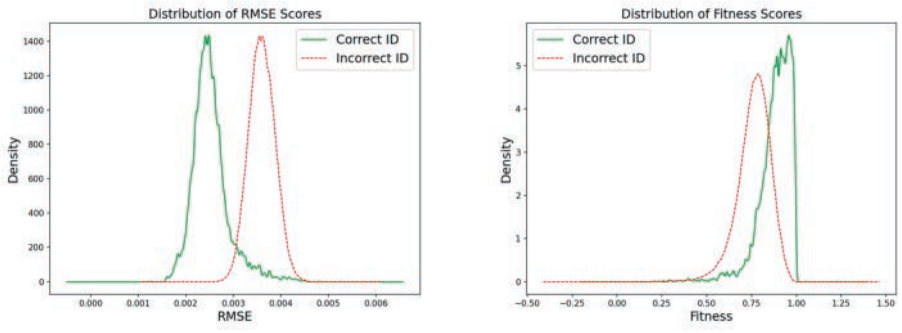


Fig. 6: ICP based recognition method

4 Results & Discussion

The metric inlier RMSE was chosen after analyzing both metrics for 1442 probes in a verification experiment. ICP was performed for all 1442 point clouds with all 160 gallery point clouds, except itself. For instance, if probes are in the gallery, ICP was performed only on 159 gallery point clouds, excluding itself. Figure 7 shows distribution plots of the



(a) Inlier RMSE

(b) Fitness

Fig. 7: Distribution plots of the cows based on the chosen metrics

same vs different cows for the fitness and inlier RMSE scores. We see that fitness is not a reliable metric as the distribution shows a considerable overlap between scores 0.75 and 1.00. However, the inlier RMSE separates the same and different cows at a score threshold of approximately 0.003. Receiver Operating Characteristic (ROC) curves are plotted for both the metrics and the result is shown in figure 8. Inlier RMSE is observed to have an Equal Error Rate (EER) of about 6.5% at a threshold of 0.0032, while Fitness has an EER of 22% at a threshold of 0.836. The results show that inlier RMSE is a better metric to classify cows in this dataset.

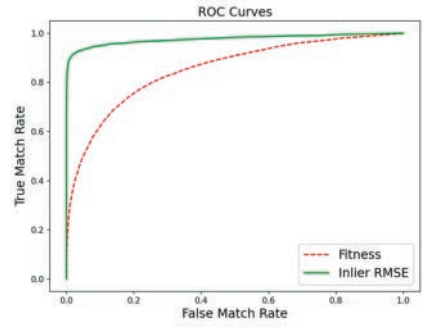


Fig. 8: ROC Curves - Fitness and Inlier RMSE

# Gallery point clouds per cow	Identification Rate (%) Min.	Max.
1	88.611	99.220
2	94.462	99.532
3	97.738	99.220
4	98.830	99.142
5	99.142	

Tab. 5: Identification Rates

In a closed set identification experiment, we see the identification rates for different number of gallery point clouds per cow. This will give us an idea of how the system could perform with respect to the amount of data in the gallery. We perform the identification experiment on 1282 probe point clouds, excluding the 160 gallery point clouds.

Table 5 shows the identification rates for different number of point clouds in gallery. Minimum and maximum identification rates refer to the extremes, where the decision is made based on the worst and the best case inlier RMSE scores respectively. In the case of 1 gallery point cloud per cow, we select the lowest (best case) and the highest (worst case) inlier RMSE scores per cow. The former yields an identification rate of 99.220% and the latter yields 88.611%. Similarly, in the cases of 2, 3 and 4 gallery point clouds per cow, we choose the average of the lowest(best cases) or highest (worst cases) 2,3 and 4 inlier RMSE scores. The average of all 5 inlier RMSE values showed an identification rate of 99.142%. The trend shows us that as we keep adding more gallery point clouds per cow we get lower RMSE scores, whose contribution is clearly reflected in the Minimum Identification Rate field.

5 Conclusion & Future Scope

The objective of this project was to investigate and prove the concept of identifying cows using their face shapes in order to improve cost efficiency and the cow's comfort. The methodology involves slightly modified existing 3D face registration and recognition methods. After acquiring face point clouds from the proposed dual 3D camera setup and registering them, ICP based recognition yields near perfect identification rate of 99.532%. The results prove that we can distinguish cows based on their face shapes and opens up further possibilities in implementing a more robust registration method, speeding up the recognition process and investigating the performance on a larger scale. While the identification rate is expected to be in a similar range, computation time will increase linearly with herd size because ICP should be performed for more cows. Most medium-sized Dutch farms have over 40 cows and a real time implementation of this system requires it to be *at least* 15 times faster (20 s per cow).

To improve the speed, implementing a faster and more accurate 3D face recognition method as explained in [Sp15] for cows on a bigger dataset would be an interesting experiment. Collecting data over a longer period of time from different types of cows will show if facial variations (natural or due to sickness) will affect the system's performance. A bigger dataset will enable further research on 3D cow face recognition using conventional and Deep Learning methods. If vision based systems out-perform the traditional electrical ones, cows will be free from IR tags around the neck.

Acknowledgement

This research was fully supported by Lely Industries N.V., located at Maassluis, The Netherlands. We would like to express our gratitude to Mr. Patrick Segeren, the Head

of Smart Components Team at Lely for initiating the project and for arranging all necessary resources for our research. We extend our thanks to the wonderful cows from Lely's test farms for their patient cooperation.

References

[Be19] Bergamini, L.; Porrello, A.; Dondona, A. C.; Del Negro, E.; Mattioli, M.; D'alterio, N.; Calderara, S.: Multi-views Embedding for Cattle Re-identification. CoRR, abs / 1902.04886, 2019.

[BM92] Besl, P. J.; McKay, N. D.: A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2):239–256, 1992.

[FB81] Fischler, M. A.; Bolles, R.C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM, 24(6):381–395, June 1981.

[JC07] Jackson, P.G.; Cockcroft, P.D.: Clinical Examination of the Head and Neck. In: Clinical Examination of Farm Animals. John Wiley & Sons, Ltd, pp. 29–50, 2007.

[Ma05] Maurer, T.; Guigonis, D.; Maslov, I.; Pesenti, B.; Tsaregorodtsev, A.; West, D.; Medioni, G.: Performance of Geometrix ActiveIDˆ TM 3D Face Recognition Engine on the FRGC Data. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops. IEEE, pp. 154–154, 2005.

[RBB09] Rusu, R. B.; Blodow, N.; Beetz, M.: Fast Point Feature Histograms (FPFH) for 3D registration. In: 2009 IEEE International Conference on Robotics and Automation. pp. 3212–3217, 2009.

[RC11] Rusu, R. B.; Cousins, S.: 3D is here: Point Cloud Library (PCL). In: 2011 IEEE International Conference on Robotics and Automation. pp. 1–4, 2011.

[Sp11] Spreeuwers, L.: Fast and Accurate 3D Face Recognition. International Journal of Computer Vision, 93(3):389–414, Jul 2011.

[Sp15] Spreeuwers, L.: Breaking the 99% barrier: optimisation of three-dimensional face recognition. IET Biometrics, 4(3):169–178, 2015.

[ZPK18] Zhou, Q.Y.; Park, J.; Koltun, V.: Open3D: A Modern Library for 3D Data Processing. arXiv:1801.09847, 2018.

BIOSIG 2020

Further Conference Contributions

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 175

Efficiency Analysis of Post-quantum-secure Face Template Protection Schemes based on Homomorphic Encryption

Jascha Kolberg[1], Pawel Drozdowski[1], Marta Gomez-Barrero[2], Christian Rathgeb[1], Christoph Busch[1]

Abstract: Since biometric characteristics are not revocable and biometric data is sensitive, privacy-preserving methods are essential to operate a biometric recognition system. More precisely, the biometric information protection standard ISO/IEC IS 24745 requires that biometric templates are stored and compared in a secure domain. Using homomorphic encryption (HE), we can ensure permanent protection since mathematical operations on the ciphertexts directly correspond to those on the plaintexts. Thus, HE allows to compute the distance between two protected templates in the encrypted domain without a degradation of biometric performance with respect to the corresponding system. In this paper, we benchmark three post-quantum-secure HE schemes, and thereby show that a face verification in the encrypted domain requires only 50 ms transaction time and a template size of 5.5 KB.

Keywords: Face Recognition, Biometric Template Protection, Post-quantum Cryptography, Homomorphic Encryption.

1 Introduction

Nowadays, biometric authentication is widely used in applications ranging from convenient smartphone unlocking to high-security border control. On the other hand, we can also observe an increase in cybercrime and databases leakages. Due to the fact that biometric characteristics are unique and cannot be changed unlike e.g. passwords, unprotected databases can be exploited to reveal enrolment data and track individuals. Hence, biometric data, amongst others, are classified as sensitive data by the European Union in the General Data Protection Regulation 2016/679 [Eu16]. Furthermore, research has proven that biometric samples can be reconstructed from unprotected templates, for instance face [Ma18], iris [Ga13], or fingerprint [Ca07]. The ISO/IEC IS 24745 standard [IS11] defines three requirements for biometric template protection (BTP): *i) unlinkability*, two protected templates cannot be linked to the same subject, *ii) renewability*, new templates can be created without the need to re-enrol and old templates can be revoked, and *iii) irreversibility*, it is impossible to retrieve original samples given only protected templates. Furthermore, the biometric performance should be preserved in the protected scheme. Therefore, BTP mechanisms are able to handle these privacy issues since templates are stored and compared in a secure domain.

[1] da/sec - Biometrics and Internet Security, University of Applied Sciences Darmstadt, Germany, {jascha.kolberg,pawel.drozdowski,christian.rathgeb,christoph.busch}@h-da.de
[2] Hochschule Ansbach, Germany, marta.gomez-barrero@hs-ansbach.de

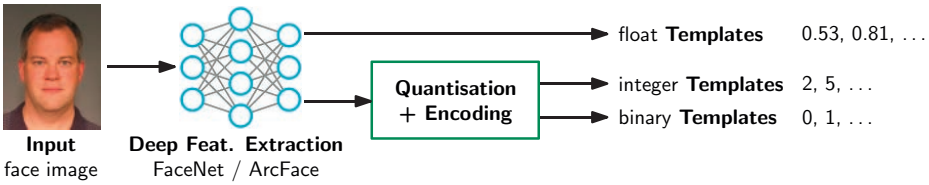


Fig. 1: Pre-processing pipeline in order to extract templates from facial input images.

By fulfilling the aforementioned requirements, the data subject's privacy is protected during the comparison as well as for leaked templates. In this context, different BTP schemes have been developed [BDL15, RU11]. One trend utilises homomorphic encryption (HE) in order to compute the biometric comparison score directly on the ciphertexts. For instance, Sadeghi *et al.* [SSW09] combine HE with garbled circuits to achieve protection in a face identification scenario. Using two HE schemes for face identification, Drozdowski *et al.* [Dr19] additionally discuss technical considerations and challenges. In order to speed up the execution for face verifications, Boddeti [Bo18] explores fully HE in combination with batching, which allows to reduce the number of homomorphic multiplications for the distance calculation. Likewise, Yasuda *et al.* [Ya15] present a specific packing method for their HE scheme to gain efficiency. Therewith, they apply template protection for different biometric modalities with feature vectors of 2,048 bit. Following the multi-modal idea, Gomez-Barrero *et al.* [Go17] propose a general framework for verification of fused modalities in the homomorphically encrypted domain.

However, a lot of publications applying HE for BTP assign the secret key to the client, thus discarding the advantages of using biometrics in general and generating a two-factor authentication system with biometrics and a secret-based knowledge. In contrast, our contribution keeps the secret key on the server side. We build upon the iris BTP scheme in [Ko19] to protect binarised face templates and benchmark recognition performance, transaction time, template size, and cryptographic security with two state-of-the-art HE BTP systems [Dr19]. The focus is on post-quantum-secure cryptography [BL17] to achieve long-term security for biometric data.

The rest of this paper is structured as follows: Section 2 describes the proposed system including the necessary pre-processing. The experimental setup and results, including the benchmark, are presented in Section 3. Finally, Section 4 concludes the paper.

2 Proposed System

2.1 Baseline system

Before we can encrypt face templates, we need to extract the features from the input images. The pre-processing pipeline for this purpose is shown in Fig. 1. Given the facial input images, two deep feature extraction algorithms, ArcFace [De19] and FaceNet [SKP15], were used to create templates of 512 floating-point values. Additionally, we applied quantisation and encoding to transform the float templates into integer and binary templates in order to be able to use additional HE schemes. Following the analysis in [Dr18], the feature

space is divided into four segments of equal size. For the integer encoding, the float values are simply mapped to the corresponding number of their sequence area. In order to have the lowest distance for adjacent areas in the binary representation, the linearly separable subcode (LSSC) [LT12] transforms each integer value into three binary digits. Float and integer templates can be compared by computing the squared Euclidean distance, while the Hamming distance is used on binary templates.

2.2 Homomorphic encryption

Homomorphic encryption schemes [Ac18] implement asymmetric cryptography with the property that specific mathematical operations on the ciphertext directly affect to the plaintext. Those additive or multiplicative homomorphic properties can generally be defined as:

$$\mathrm{Enc}\,(A+B) = \mathrm{Enc}\,(A) \diamond \mathrm{Enc}\,(B) \tag{1}$$

$$\mathrm{Enc}\,(A \cdot B) = \mathrm{Enc}\,(A) \circ \mathrm{Enc}\,(B) \tag{2}$$

Hence, we have an operation $\diamond$ that results in the sum of two plaintexts when it is applied to both corresponding ciphertexts. Another operation $\circ$ is used to achieve a multiplication. The specific operations depend on the selected HE scheme. Moreover, not all HE schemes support all operations. Therefore, depending on the used biometric templates and the required distance computations, different HE schemes [Ac18] should be utilised.

With the focus on post-quantum-security [BL17], the following three crypto schemes are selected. In order to compute the squared Euclidean distance in the encrypted domain, two HE schemes are utilised. The encryption scheme by Cheon-Kim-Kim-Song (CKKS) [Ch17] supports homomorphic operations on floating point templates and the Brakerski/Fan-Vercauteren (BFV) [FV12] scheme is applied on the integer templates. On the other hand, the computation of the Hamming distance in the encrypted domain can efficiently be done with the N-th degree truncated polynomial ring (NTRU) [HPS98] scheme, if the parameters are selected in a way that the decryption automatically performs a modulo-2 operation. Then one addition directly results in the *XOR* of probe and reference. The security of all three schemes is based on the ring-learning-with-errors problem, which, using a quantum algorithm, can be reduced to the shortest vector problem over ideal lattices [LPR10]. Thus, granting long-term security for biometric templates.

2.3 Protected system

Based on the aforementioned considerations, we can build our HE protected system. Since we are not interested in a two-factor authentication, where the secret key is assigned to the client, the secret key needs to be stored at server-side. However, placing the decryption key next to the encrypted templates in the database server (DB) threatens the whole purpose. Hence, we need an additional authentication server (AS) in our infrastructure, which works as a trusted third party. The structure of this system and a verification transaction are illustrated in Fig. 2.

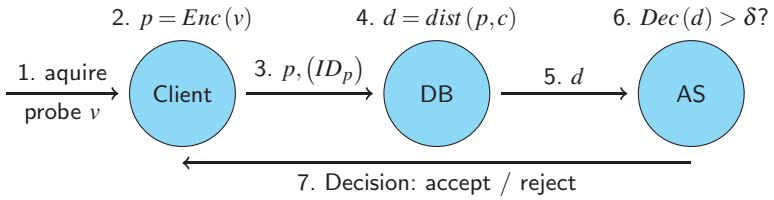


Fig. 2: Verification transaction of the BTP system using homomorphic encryption.

1. The client captures the biometric characteristics and pre-processes the data, resulting in a probe feature vector v.
2. The client encrypts v with the public key to get the protected probe p.
3. The encrypted probe p is sent to the database server (DB). In a verification scenario, the client additionally transfers an ID claim.
4. DB computes the distance d between probe p and reference(s) c_i in the encrypted domain.
5. This encrypted distance d is forwarded to the authentication server (AS).
6. AS decrypts d using the secret key and compares the result with a decision threshold. Alternatively, AS could also sort all computed distances in identification mode.
7. The final accept/reject decision is revealed to the client.

This architecture assumes the honest-but-curious model, where the parties stick to the protocol, but may try to learn as much information as possible. This implies that DB and AS do not collude in order to decrypt the database or incoming probes. The client encrypts its probe before sending it to the DB, which only operates on encrypted templates to compute the encrypted distance. The AS, which possesses the secret decryption key, receives only protected distance values and thus does not learn sensitive information from neither the probe nor the reference. The transmission channel between parties can additionally be protected by TLS. For higher privacy, the decision (in 7.) could be returned to the DB, which forwards it to the client in order to conceal the identity of the client device from AS.

3 Experimental Evaluation

3.1 Experimental Setup

The experiments were run on a frontal image subset of the FERET database [Ph00] comprising 6,963 samples of 563 subjects. Both feature extraction methods use their freely available pre-trained model, which allows for reproducibility of our research. Furthermore, our BTP systems are implemented based on the open-source crypto implementations in the Microsoft SEAL HE library[3] (CKKS and BFV) and the NTRU iris template protection system [Ko19].

[3] `https://github.com/Microsoft/SEAL`

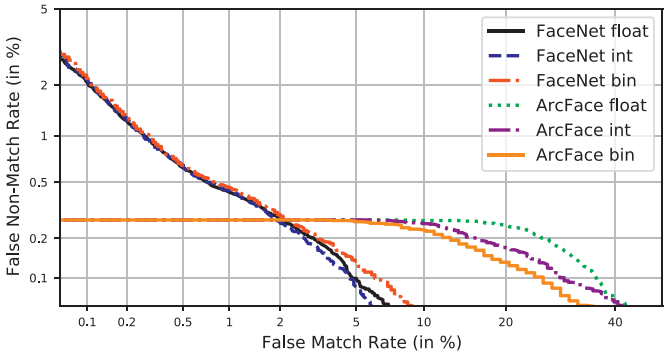


Fig. 3: DET plot showing the false match rates and false non-match rates for different template types in the verification scenario. The performance is identical for unprotected and protected systems.

Tab. 1: Rank-1 identification rates in % for different feature types.

Rank-1	float	integer	binary
ArcFace (%)	99.03	98.98	99.03
FaceNet (%)	98.50	98.42	98.36

In order to compare the biometric performance of different template representations, verification scores as well as rank-1 mated identification scores are computed. For the genuine verification, all mated samples are compared and for the impostor scores, only the first sample of each subject is compared with the first sample of all other subjects. Furthermore, the rank-1 identification scores include all mated comparisons.

The timing of relevant functions was conducted within a virtualised (single-core) Linux on a commodity notebook running an Intel Core i7 2.7 GHz CPU and 16 GB DDR4 RAM. While the SEAL library is written in C++, the Python3 code of the NTRU implementation is executed with PyPy3[4] for an additional speed-up.

3.2 Results

Biometric performance evaluation. Fig. 3 shows that biometric verification performance is preserved across the different template types. Especially for a false match rate below 2%, the DET curves for all types are almost identical. However, looking at the rank-1 identification rates of all mated comparisons in Tab. 1 reveals minor variations for the different template types due to the quantisations. In general we can conclude, that the biometric performance remains stable across different face template representations.

Execution time and file size. Generally, we could observe that execution times and file sizes slightly increase for higher security levels. However, the relative speed-up between different template representations stayed the same and hence only results for a security level of 128 bits are presented.

[4] `https://pypy.org`

Tab. 2: Median execution time and standard deviation of relevant functions. The comparison includes distance computation, decryption, and deriving the final decision.

128 bits Security	CKKS (float)	BFV (int)	NTRU (bin)
Key generation (ms)	779 (±4)	255 (±5)	362 (±84)
Encryption (ms)	6 (±2)	76 (±1)	27 (±5)
Comparison (ms)	3391 (±10)	618 (±26)	23 (±3)

Tab. 3: File size of keys and a single template for the evaluated encryption schemes.

128 bits Security	CKKS (float)	BFV (int)	NTRU (bin)
Keys	99 MB	12 MB	6 KB
Template	516 KB	132 KB	5.5 KB

The execution times for relevant functions are depicted in Tab. 2. The key generation of all encryption schemes is done in less than one second and is negligible since it is a one time effort at enrolment. The encryption times refer to a single template and need to be multiplied with the number of references in the database during the system setup. An encryption takes about 6 ms for float features, 76 ms for integer features, and 27 ms for binary features. Analogously to the encryption, we need to multiply the comparison times with the number of references for each performed identification. A comparison in BFV (618 ms) is five times faster than CKKS (3,391 ms) and 25 times slower than NTRU (23 ms). Additionally, for both identification and verification, the probe needs to be encrypted before the comparison. Thus, this single encryption affects the verification much more than the identification scenario. It occurs that the CKKS encryption is faster than the other encryptions, while it needs much longer on other operations. The NTRU encryption relies on new random polynomials for each block and their generation takes apparently more time than its corresponding function within CKKS. Comparing the file sizes, as shown in Tab. 3, reveals intuitive results. The keys of CKKS are with 99 MB the biggest, while BFV requires 12 MB and NTRU 6 KB. The same order holds for the template sizes; 516 KB for CKKS, 132 KB for BFV, and 5.5 KB for NTRU.

Simulating a company database with 1,000 employees would require a database storage of around 500 MB for CKKS, 130 MB for BFV, and 6 MB for NTRU. Using the same system in identification mode would take around one hour in the CKKS scheme, nearly 12 minutes in the BFV scheme, and 50 seconds with NTRU.

Security analysis. Finally, all three HE schemes are based on ideal lattices and hence found to be post-quantum-secure [BL17], which grants us *irreversibility*. The encryption functions utilise a random factor with the effect that encrypting the identical plaintext twice, results in two *unlinkable* ciphertexts. *Renewability* can be achieved by exchanging the key pair and re-encrypting the database. Since clients only operate with the public key, no re-enrolment is required.

Summary. These results show that real time verifications in the encrypted domain are possible for integer and binary face templates. However, when it comes to identification, only

the comparison of binary templates is fast enough to support a reasonable transaction time. As demonstrated, the biometric performance remains largely unaffected for all presented template representations.

4 Conclusions

This work showed that the most important requirement for efficient template protection, transforming float templates into integer or binary templates, has negligible impact on the biometric recognition accuracy. Furthermore, the three ISO/IEC IS 24745 requirements *irreversibility*, *unlinkability*, and *renewability* are fulfilled by the evaluated HE schemes, CKKS, BFV, and NTRU. Additional long-term template security is granted by their post-quantum-secure design. Since we worked with a public database and only used open-source software, all results from this paper are reproducible. Most importantly, using binary face templates, a verification in the encrypted domain is done within 50 ms on an ordinary notebook, which also allows to apply NTRU HE in limited identification scenarios. Those results demonstrate the practicability of biometric template protection for face verification even on off-the-shelf hardware. Future work will evaluate efficient biometric identification in the homomorphic domain including computational workload reduction methods [DRB19].

Acknowledgements

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

References

[Ac18] Acar, A.; Aksu, H.; Uluagac, A. S.; Conti, M.: A Survey on Homomorphic Encryption Schemes: Theory and Implementation. ACM Computing Surveys, 51(4):1–35, 2018.

[BDL15] Barni, M.; Droandi, G.; Lazzeretti, R.: Privacy Protection in Biometric-Based Recognition Systems: A Marriage Between Cryptography and Signal Processing. IEEE Signal Processing Magazine, 32(5):66–76, 2015.

[BL17] Bernstein, D. J.; Lange, T.: Post-quantum cryptography. Nature, 549(7671):188–194, 2017.

[Bo18] Boddeti, V. N.: Secure Face Matching Using Fully Homomorphic Encryption. In: Proc. of Int. Conf. on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–10, 2018.

[Ca07] Cappelli, R.; Maio, D.; Lumini, A.; Maltoni, D.: Fingerprint Image Reconstruction From Standard Templates. Trans. on Pattern Analysis and Machine Intelligence, 29(9), 2007.

[Ch17] Cheon, J. H.; Kim, A.; Kim, M.; Song, Y.: Homomorphic Encryption for Arithmetic of Approximate Numbers. In: Proc. Asiacrypt. Springer, pp. 409–437, 2017.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proc. CVPR. pp. 4690–4699, 2019.

[Dr18] Drozdowski, P.; Struck, F.; Rathgeb, C.; Busch, C.: Benchmarking Binarisation Schemes for Deep Face Templates. In: Int. Conf. on Image Processing (ICIP). IEEE, pp. 1–5, 2018.

[Dr19] Drozdowski, P.; Buchmann, N.; Rathgeb, C.; Margraf, M.; Busch, C.: On the Application of Homomorphic Encryption to Face Identification. In: Proc. BIOSIG. pp. 1–8, 2019.

[DRB19] Drozdowski, P.; Rathgeb, C.; Busch, C.: Computational Workload in Biometric Identification Systems: An Overview. IET Biometrics, 8(6):351–368, 2019.

[Eu16] European Parliament: . EU Regulation 2016/679 (General Data Protection Regulation), 2016.

[FV12] Fan, J.; Vercauteren, F.: Somewhat Practical Fully Homomorphic Encryption. IACR Cryptology ePrint Archive, 2012:144, 2012.

[Ga13] Galbally, J.; Ross, A.; Gomez-Barrero, M.; Fierrez, J.; Ortega-Garcia, J.: Iris Image Reconstruction From Binary Templates: An Efficient Probabilistic Approach Based on Genetic Algorithms. Computer Vision and Image Understanding, 117(10):1512–1525, 2013.

[Go17] Gomez-Barrero, M.; Maiorana, E.; Galbally, J.; Campisi, P.; Fierrez, J.: Multi-Biometric Template Protection Based on Homomorphic Encryption. Pattern Recognition, 67:149–163, 2017.

[HPS98] Hoffstein, J.; Pipher, J.; Silverman, J. H.: NTRU: A Ring-Based Public Key Cryptosystem. In: Int. Algorithmic Number Theory Symposium. Springer, pp. 267–288, 1998.

[IS11] ISO/IEC JTC1 SC27 Security Techniques: . ISO/IEC 24745:2011. Information Technology - Security Techniques - Biometric Information Protection. ISO/IEC, 2011.

[Ko19] Kolberg, J.; Bauspieß, P.; Gomez-Barrero, M.; Rathgeb, C.; Dürmuth, M.; Busch, C.: Template Protection based on Homomorphic Encryption: Computationally Efficient Application to Iris-Biometric Verification and Identification. In: Proc. WIFS. 2019.

[LPR10] Lyubashevsky, V.; Peikert, C.; Regev, O.: On Ideal Lattices and Learning with Errors Over Rings. In: Annual Int. Conf. on the Theory and Applications of Cryptographic Techniques. Springer, pp. 1–23, 2010.

[LT12] Lim, M. H.; Teoh, A. B. J.: A Novel Encoding Scheme for Effective Biometric Discretization: Linearly Separable Subcode. IEEE Trans. on Pattern Analysis and Machine Intelligence, 35(2):300–313, 2012.

[Ma18] Mai, G.; Cao, K.; Yuen, P. C.; Jain, A. K.: On the Reconstruction of Face Images from Deep Face Templates. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2018.

[Ph00] Phillips, P. J.; Moon, H.; Rizvi, S. A.; Rauss, P. J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(10):1090–1104, 2000.

[RU11] Rathgeb, C.; Uhl, A.: A Survey on Biometric Cryptosystems and Cancelable Biometrics. EURASIP Journal on Information Security, 2011(1):3, 2011.

[SKP15] Schroff, F.; Kalenichenko, D.; Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: Proc. CVPR. pp. 815–823, 2015.

[SSW09] Sadeghi, A. R.; Schneider, T.; Wehrenberg, I.: Efficient Privacy-preserving Face Recognition. In: Int. Conf. on Information Security and Cryptology. Springer, pp. 229–244, 2009.

[Ya15] Yasuda, M.; Shimoyama, T.; Kogure, J.; Yokoyama, K.; Koshiba, T.: New Packing Method in Somewhat Homomorphic Encryption and Its Applications. Security and Communication Networks, 8(13):2194–2213, 2015.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb, K. Raja and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 183

A robust fingerprint presentation attack detection method against unseen attacks through adversarial learning

Joao Afonso Pereira[1], Ana F. Sequeira[1], Diogo Pernes[1,2], Jaime S. Cardoso[1,3]

Abstract: Fingerprint presentation attack detection (PAD) methods present a stunning performance in current literature. However, the *fingerprint PAD generalisation problem* is still an open challenge requiring the development of methods able to cope with sophisticated and unseen attacks as our eventual intruders become more capable. This work addresses this problem by applying a regularisation technique based on an adversarial training and representation learning specifically designed to to improve the PAD generalisation capacity of the model to an unseen attack. In the adopted approach, the model jointly learns the representation and the classifier from the data, while explicitly imposing invariance in the high-level representations regarding the type of attacks for a robust PAD. The application of the adversarial training methodology is evaluated in two different scenarios: i) a handcrafted feature extraction method combined with a Multilayer Perceptron (MLP); and ii) an end-to-end solution using a Convolutional Neural Network (CNN). The experimental results demonstrated that the adopted regularisation strategies equipped the neural networks with increased PAD robustness. The adversarial approach particularly improved the CNN models' capacity for attacks detection in the unseen-attack scenario, showing remarkable improved APCER error rates when compared to state-of-the-art methods in similar conditions.

Keywords: Fingerprint presentation attack detection, adversarial learning, transfer learning.

1 Introduction

Biometric recognition is nowadays a mature technology used in many government and civilian applications such as e-passports, ID cards, border control and in most of unlock/authentication systems present in handheld devices. Fingerprint recognition systems (FRS) in particular are widely used probably having been this the first biometric trait used to identify people. Fingerprint presentation attack detection (FPAD) methods have been developed as an attempt to overcome the vulnerability of FRS to spoofing. However, most of the traditional approaches have been quite optimistic about the behavior of the intruder, assuming the use of a previously known type of attack sample. This assumption has led to the overestimation of the performance of the methods, using both live and spoof samples to train the predictive models and evaluate each type of fake samples individually [SC15].

The presentation attack detection (PAD) generalisation capacity of a model to unseen attacks, has been addressed before regarding iris, fingerprint and face. However, it still remains a challenging topic. Whether in research or deployment of PAD systems in commercial applications, if the classification models are designed and evaluated using bona fide presentations and presentation attack instruments (PAI) belonging only to specific species

[1] INESC TEC, Porto, Portugal, Email: {joao.a.pereira,ana.f.sequeira, diogo.pernes, jaime.cardoso}@inesctec.pt
[2] Faculdade de Ciencias da Universidade do Porto, Porto, Portugal
[3] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

(PAISp), then the case when the model is confronted with a PAISp which is significantly different from the ones used for training is overlooked. In the worst case scenario, such sample may have higher probability to circumvent the system than the ones drawn from the original training dataset and the model may fail to generalise robustly and detect the threat. To solve this research question is necessary to develop robust methods to cope with sophisticated and unseen attacks as our eventual intruders become more capable and successfully develop new spoofing techniques.

The pioneer work in the evaluation of PAD methods across different types and unseen PAISp appeared in the fingerprint domain with the work of Marasco and Sansone [MS11]. The works of Rattani & Ross [RSR15] and Sequeira & Cardoso [SC15], despite using different approaches, both relied on the idea of enforcing the knowledge of the bona fide (BF) presentations over the presentation attack (PA) to better deal with unseen PAISp. With the rise of deep learning (DL) techniques, PAD methods based on deep representations were proposed using the binary approach [Me15, Pi18]. Followed by works tackling DL fingerprint PAD methods robustness to unseen PAISp. In [To18], was proposed a highly accurate method based on CNNs and own multi-spectral short wave infrared imaging. The LivDet competition series in 2015 [Mu15] included evaluation with unseen attacks, however unfortunately this scenario was not tested in following editions. The PAD generalisation problem has been addressed regarding other biometric traits. Regarding iris, Sequeira *et al* [Se16] stated that whenever a new PAISp is presented in the test step, the performance of the classifier drops and improvement can be obtained using BF one-class classification; and in [Fe19a] a successful adversarial strategy is proposed. Nevertheless, most of the recent approaches, either make overly optimistic assumptions about the attacker - binary classification approaches - or only use part of the data (and therefore, of the knowledge) available at training time to design the models - one-class approaches. Alternatively, the approach evaluated in this work uses the information of both BF and known PA and is robust to unseen PAI species.

In this work, the *FPAD generalisation problem* is addressed by means of a regularisation technique applied to artificial neural networks combining adversarial training with representation learning. In this approach, designed to improve the generalisation capacity to unseen attacks, the proposed model jointly learns the representation and the classifier from the data, while explicitly imposing 'PAI-species'-invariance in the high-level representations for a robust PAD method. The algorithm applied here was presented by Ferreira *et al* [Fe19b] in the context of sign language recognition, with a later application to iris PAD [Fe19a]. This approach builds on those initially introduced by Ganin *et al* [GL15], for domain adaptation, and Feutry *et al* [Fe18], to learn anonymized representations.

The contributions of this work are then two-fold: 1) the application of the adversarial training concept to the generalisation to unseen attacks problem in fingerprint PAD; and 2) the evaluation of the adversarial training methodology in two different scenarios: i) a handcrafted feature extraction method combined with a Multilayer Perceptron (MLP); ii) an end-to-end solution using a Convolutional Neural Network (CNN).

The main definitions related to PAD concepts used throughout this paper are the ones stated in the International Standard ISO/IEC 30107-3 Information Technology — Biometric presentation attack detection — Part 3: Testing and reporting [IS17].

This paper is organised as follows. This section summarises the related and proposed work and how it addresses the research question posed. In section 2 the methodology used is detailed. Section 3 describes the experimental setup including the results and discussion. Section 4 concludes the work with the final remarks.

2 Methodology

This section summarises the methodology from Ferreira *et al* [Fe19a], which is adopted here with the appropriate adjustments. The underlying idea behind this approach is that, in order to generalise well to unseen attacks, the model should not specialize in discriminating any of the PAI species (PAISp) presented at training time and, therefore, the learned internal representations should be invariant to the PAISp. For this purpose, the model combines an adversarial approach with a species-transfer training objective, which are described in the remaining of the section. The high-level architecture of the model is summarized in Figure 1. Throughout this section, it should be assumed that one has access to a labeled dataset $\mathbb{X} = \{\boldsymbol{X}_i, y_i, s_i\}_{i=1}^{N}$ of N samples, where $\boldsymbol{X}_i$ represents the i-th input sample, and y_i and s_i denote the corresponding class label (*bona fide* or *attack*) and the PAI species (only defined for attack samples), respectively. Let $\mathbb{X}^{bf}$ and $\mathbb{X}^{a}$ be these partitions of $\mathbb{X}$ for bona-fide and attack samples, respectively, and N^{bf} and N^{a} their respective cardinality.

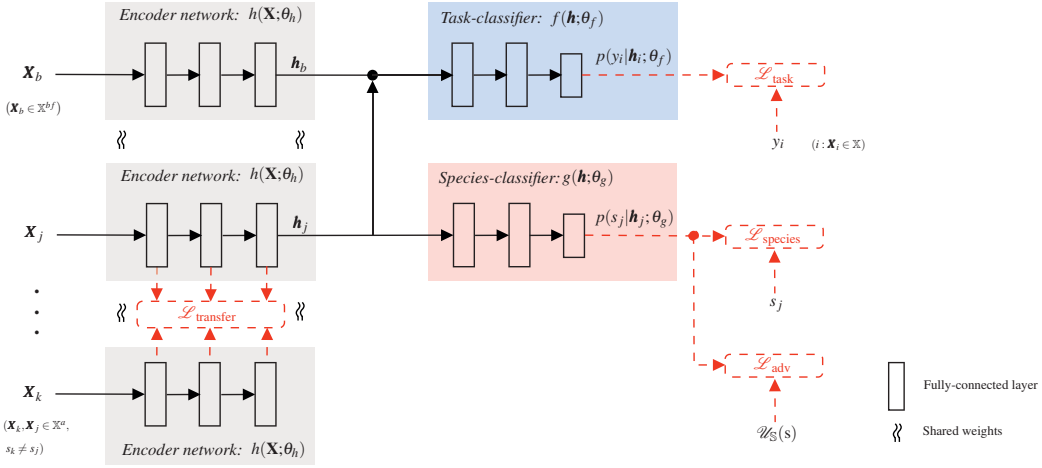


Fig. 1: The architecture of the proposed species-invariant neural network (from [Fe19a]).

2.1 Adversarial learning

The model comprises three main sub-networks: (i) an encoder network $h(\cdot;\theta_h)$ that receives input samples and maps them to a latent space; (ii) a *task-classifier* network $f(\cdot;\theta_f)$ which aims to distinguish attack and bona fide samples, mapping latent representations to the corresponding class probabilities; and (iii) a *species-classifier* network $g(\cdot;\theta_g)$ that receives latent representations from attack samples and aims to predict the corresponding

PAI species. In order to learn 'PAI-species'-invariant latent representations, an adversarial learning scheme is adopted. The species-classifier is trained to minimize the classification loss of the PAI-species:

$$\min_{\theta_g} \mathscr{L}_{\text{species}}(\theta_h, \theta_g) = \min_{\theta_g} \left\{ -\frac{1}{N^a} \sum_{i=1}^{N^a} \log p(s_i | h(\boldsymbol{X}_i; \theta_h); \theta_g) \right\}, \ \boldsymbol{X}_i \in \mathbb{X}^a. \tag{1}$$

Simultaneously, the task-classifier and the encoder are jointly trained to minimize the classification loss between attacks and bona fide samples, while trying to keep the PAI-species classification close to random guessing (i.e., close to a uniform distribution):

$$\min_{\theta_h, \theta_f} \mathscr{L}(\theta_h, \theta_f, \theta_g) = \min_{\theta_h, \theta_f} \left\{ \mathscr{L}_{\text{task}}(\theta_h, \theta_f) + \lambda \mathscr{L}_{\text{adv}}(\theta_h, \theta_g) \right\}, \tag{2}$$

where

$$\mathscr{L}_{\text{task}}(\theta_h, \theta_f) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i | h(\boldsymbol{X}_i; \theta_h); \theta_f), \tag{3}$$

$$\mathscr{L}_{\text{adv}}(\theta_h, \theta_g) = \frac{1}{N^a} \sum_{i=1}^{N^a} D_{\text{KL}}(\mathscr{U}_{\mathbb{S}}(\text{s}) || p(\text{s} | h(\boldsymbol{X}_i; \theta_h); \theta_g), \ \boldsymbol{X}_i \in \mathbb{X}^a. \tag{4}$$

Here, $\mathscr{U}_{\mathbb{S}}$ denotes a uniform distribution over the set of PAI-species present in the training set.

2.2 Species-transfer objective

In addition to the adversarial training, a species-transfer objective is employed to further encourage the latent representations to be species-invariant. This objective enforces the means of the latent representations of different species to coincide. Therefore, this is a weaker constraint than the one imposed by the adversarial objective, but it has a beneficial effect by speeding up the convergence to invariant representations.

Specifically, a layer-wise loss $\mathscr{D}^{(m)}$ between the hidden representations $h^{(m)}(\cdot; \theta_h)$ of two species s and t at the output of the m-th layer of the encoder is defined as:

$$\mathscr{D}^{(m)}(s, t; \theta_h) = \left|\left| \frac{1}{N_s} \sum_{i:\ s_i = s} h^{(m)}(\boldsymbol{X}_i; \theta_h) - \frac{1}{N_t} \sum_{j:\ s_j = t} h^{(m)}(\boldsymbol{X}_j; \theta_h) \right|\right|_2^2, \tag{5}$$

where $||\cdot||_2$ is the ℓ^2-norm, and N_s and N_t denote the number of training examples of species s and t, respectively. The overall species-transfer loss $\mathscr{L}_{\text{transfer}}$ is then a weighted sum of the losses computed at each layer of the *encoder* network:

$$\mathscr{L}_{\text{transfer}}(\theta_h) = \sum_{m=1}^{M} \beta^{(m)} \mathscr{L}^{(m)}_{\text{transfer}}(\theta_h) = \sum_{m=1}^{M} \beta^{(m)} \sum_{s \in \mathbb{S}} \sum_{\substack{t \in \mathbb{S}, \\ t \neq s}} \mathscr{D}^{(m)}(s, t; \theta_h), \tag{6}$$

where $\beta^{(m)} \geq 0$ is a hyperparameter that controls the relative importance of the loss obtained at the m-th layer and the species-transfer loss at the m-th layer is the sum of the pairwise distances between all PAISp.

The overall objective function of the encoder and task classifier is then the combination of equations (2) and (6):

$$\min_{\theta_h,\theta_f} \mathscr{L}(\theta_h,\theta_f,\theta_g) = \min_{\theta_h,\theta_f} \left\{ \mathscr{L}_{\text{task}}(\theta_h,\theta_f) + \lambda \mathscr{L}_{\text{adv}}(\theta_h,\theta_g) + \gamma \mathscr{L}_{\text{transfer}}(\theta_h) \right\}, \quad (7)$$

where $\gamma \geq 0$ is the weight that controls the relative importance of the species-transfer term. The objective for the species-classifier remains unchanged, i.e. as in equation (1).

3 Experimental setup

PAD Performance Evaluation Metrics: The *Attack Presentation Classification Error Rate* (APCER) and the *Bona-fide Presentation Classification Error Rate* (BPCER) for an APCER of 5% (*BPCER@APCER=5%*) as defined in the ISO/IEC 30107-3 [IS17]. The Equal Error Rate (EER) analyses the distributions of the APCER and BPCER and corresponds to the minimum value where both are equal.

Dataset: The Fingerprint Liveness Detection Competition 2015 (LivDet2015) [Mu15] training dataset comprises a set of five subsets, each one corresponding to a specific fingerprint sensor. For each sensor there are bona fide samples and different types of PAI.

Evaluation protocols: The adopted framework is denominated "unseen-attack", as the presentation attack instrument seen in the testing phase is unknown to the model. Thus, the methods are evaluated by saving one type of attack - PAI species - for testing while the training is done with the remaining presentation attack instruments and bona fide samples.

Handcrafted feature extraction method: The extracted features that served as input for the MLP were the histogram of intensity, the histogram of the Local Binary Patterns (LBP) [OPM02] and the histogram of the Local Phase Quantization [OH08].

Implementation details: The models were implemented in Python with the PyTorch library. The training phase was conducted with the Adam optimizer and a batch size of 16. The learning rate and the ℓ^2 regularization weight were both set to $1e^{-04}$. The hyperparameters λ and γ, specific to the adopted regularization, were optimized through a grid search and cross-validation on the training dataset, varying on logaritmic scale in the interval [$1e^{-03}$,1]. The $\mathscr{L}_{\text{transfer}}$ penalty is applied to the last layer of the encoder network. Regarding the architecture of the MLP, the encoder corresponds to [(FC(128) →ReLU) x 3] and the classifiers also to [(FC(128) →ReLU) x 3], where FC(n) notes a fully connected layer with n neurons. For the CNN model, the encoder corresponds to [(C(64) →ReLU) x 2 →MaxPool →(C(128) →ReLU) x 2 →MaxPool →(C(256) →ReLU) x 4], where C(f) notes a convolutional layer with f filters, kernel 3x3, stride 1 and padding 1. The CNN's classifiers both correspond to [(FC(4096) →ReLU) x 2 →(FC(1000) →ReLU)].

Results and discussion: In Table 2, the results of the baseline methods (*MLP* and *CNN*) and their respective regularised versions (MLP_{reg} and CNN_{reg}) are displayed. Comparing

the performance of the baseline and regularised versions, it can be observed that: i) regarding the MLP, except for the Hi Scan sensor, in all the cases there is a significant improvement in at least 2 out of the 3 presented metrics; and ii) regarding the CNN, there is a significant improvement without exception in all error rates, with a particular significant improvement of the APCER value from 4.12% to 0.81% (for the average of the five sensors). From these observations, it can be stated with confidence that, overall, the regularisation technique improves the PAD robustness of both the models.

Still, it is arguable that the performance of the MLP, even the baseline version, outperforms the CNN results. Nevertheless, it should be noted that: i) the first scenario is taking advantage of rich handcrafted features; and ii) the data available for training is not enough to take the best out of the CNN learning capabilities. Thus, on the one hand the end-to-end solution provided by the CNN saves a considerable effort in the computation of the feature extraction step and, on the other hand, increasing the amount of training data will certainly increase the performance of these models, as there is a high potential for growth.

Tab. 1: Baseline and proposed regularised approaches - Cross Match, Digital Persona and Green Bit sensors. ($BPCER@APCER = 5\%$ noted by $BPCER@5$.)

Method	**PAD metrics (%)**								
	Cross Match			*Digital Persona*			*GreenBit*		
	APCER	BPCER@5	EER	APCER	BPCER@5	EER	APCER	BPCER@5	EER
MLP	**0.07**	7.57	4.33	0.00	0.53	0.45	**0.70**	**0.20**	1.10
MLPreg	0.13	**4.30**	**3.70**	**0.00**	**0.00**	**0.30**	**0.70**	0.63	**0.93**
CNN	5.00	6.25	8.70	5.60	10.80	7.28	3.03	14.13	7.05
CNNreg	**1.07**	**4.65**	**2.82**	**0.60**	**3.85**	**2.45**	**0.60**	**2.93**	**1.63**

Tab. 2: Baseline and proposed regularised approaches - Hi Scan and Time Series sensors, as well as the average of the results for the 5 sensors. ($BPCER@APCER = 5\%$ noted by $BPCER@5$.)

Method	**PAD metrics (%)**								
	Hi Scan			*Time Series*			Average of the 5 sensors		
	APCER	BPCER@5	EER	APCER	BPCER@5	EER	APCER	BPCER@5	EER
MLP	**0.30**	**2.83**	**3.03**	**0.00**	**0.03**	0.60	**0.21**	2.23	1.90
MLPreg	1.30	3.60	3.38	**0.00**	**0.03**	**0.10**	0.43	**1.71**	**1.68**
CNN	5.60	20.15	11.25	1.37	9.10	4.07	4.12	12.09	7.67
CNNreg	**1.20**	**1.21**	**1.04**	**0.60**	**6.30**	**2.70**	**0.81**	**3.79**	**2.13**

Despite the evidences showed in favour of the effectiveness of the regularisation technique, it is crucial to compare the results obtained with the proposed approach against the current state-of-the-art DL based PAD that tackle the unseen-attack scenario. This is not an easy task as most works still opt for a more traditional approach, based on binary classification limited to one type of attack at a time. From the available literature using similar databases and addressing the generalisation problem, stands out the meritory initiative of Fingerprint LivDet2015 of evaluating the methods with some unseen types of PAISp.

Table 3 presents the results of the proposed regularised CNN version, *CNNreg*, alongside with the comparable literature methods currently available. The comparison is made with the best results from the LivDet2015 [Gh17, Mu15] for common subsets of the used database, as well as with an additional recent publication [Pa19]. From the observed re-

sults, it is remarked the significant improvement of the *CNNreg* in two out of three sensors and undoubtedly when considering the average values. In particular, the *CNNreg* provided an APCER value of 0.76% against 2.09% and 6.33% of the other methods (for the average of the three sensors).

Tab. 3: Literature and proposed approach.($BPCER@APCER = 5\%$ noted by $BPCER@5$.)

Method	**PAD metrics (%)**							
	Cross Match		*Digital Persona*		*GreenBit*		Average	
	APCER	BPCER@5	APCER	BPCER@5	APCER	BPCER@5	APCER	BPCER@5
Proposed CNNreg	1.07	4.65	**0.60**	**3.85**	**0.60**	**2.93**	**0.76**	**3.81**
LivDet2015 [Gh17, Mu15]	1.68	$\approx$ **0.80**	**0.60**	$\approx$ 10.00	4.00	$\approx$ 5.00	2.09	$\approx$ 5.27
Park et al [Pa19]	**0.00**	-	11.00	-	8.00	-	6.33	-

4 Conclusions

This work addresses the *fingerprint PAD generalisation problem* through an adversarial training objective which combines representation learning and artificial neural networks. The method is specifically designed to address the generalisation capacity to an unseen attack by enforcing the learning of the task of distinguishing the bona fide from the attack presentations while ensuring the invariance between the different type of the PAI species.

Comparing the baseline and regularised versions, it can be stated that, overall, the regularisation technique improves the PAD robustness of both the models. Despite the fact that the *MLPreg* fed with rich handcrafted features proved to be competitive, the fact is that *CNNreg* has more potential for growth and for increasing its performance in the future.

The comparison of the proposed approach against the current DL based PAD methods that tackle the unseen-attack scenario, is not an easy task as most works still opt for a more traditional approach based on binary classification limited to one type of attack at a time. Still, from the comparison with the available literature using similar databases and addressing the generalisation problem, it is verified a significant superiority of the *CNNreg* in two out of three sensors and undoubtedly when considering the average values.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020 and by Ph.D. Grant SFRH/BD/129600/2017.

References

[Fe18] Feutry, Clément; Piantanida, Pablo; Bengio, Yoshua; Duhamel, Pierre: Learning anonymized representations with adversarial neural networks. arXiv:1802.09386, 2018.

[Fe19a] Ferreira, Pedro; Sequeira, Ana Filipa; Pernes, Diogo; Rebelo, Ana; Cardoso, Jaime S.: Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations. In: Proceedings of the 18th BIOSIG. 2019.

[Fe19b] Ferreira, Pedro M.; Pernes, Diogo; Rebelo, Ana; Cardoso, Jaime S.: Learning signer invariant representations with adversarial training. In: 12th ICMV. 2019.

[Gh17] Ghiani, Luca; Yambay, David A.; Mura, Valerio; Marcialis, Gian Luca; Roli, Fabio; Schuckers, Stephanie A.: Review of the Fingerprint Liveness Detection (LivDet) competition series: 2009 to 2015. Image and Vision Computing, 58:110 – 128, 2017.

[GL15] Ganin, Yaroslav; Lempitsky, Victor: Unsupervised Domain Adaptation by Backpropagation. In: Proc. 32nd Int. Conf. ML. volume 37, Lille, France, pp. 1180–1189, 2015.

[IS17] ISO/IEC JTC1 SC37: Information Technology - Biometrics - Presentation attack detection Part 3: Testing and Reporting. ISO Int. Organization for Standardization, 2017.

[Me15] Menotti, D.; Chiachia, G.; Pinto, A.; Robson Schwartz, W.; Pedrini, H.; Xavier Falcao, A.; Rocha, A.: DeepRep.Iris,Face,and Fingerp.Spoof.Det. TIFS, 10(4):864–879, 2015.

[MS11] Marasco, Emanuela; Sansone, Carlo: On the Robustness of Fingerprint Liveness Detect. Alg. against New Materials used for Spoofing. In: BIOSIGNALS. pp. 553–558, 2011.

[Mu15] Mura, Valerio; Ghiani, Luca; Marcialis, Gian; Roli, Fabio; Yambay, David; Schuckers; Schuckers, Stephanie: LivDet2015-Fingerprint Liveness Detect. Competition. 09 2015.

[OH08] Ojansivu, Ville; Heikkilä, Janne: Blur Insensitive Texture Classification Using Local Phase Quantization. In (Elmoataz, Abderrahim; Lezoray, Olivier; Nouboud, Fathallah; Mammass, Driss, eds): Image and Signal Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 236–243, 2008.

[OPM02] Ojala, T.; Pietikainen, M.; Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):971–987, July 2002.

[Pa19] Park, E.; Cui, X.; Nguyen, T. H. B.; Kim, H.: Presentation Attack Detection Using a Tiny Fully Convolutional Network. IEEE Transactions on Information Forensics and Security, 14(11):3016–3025, 2019.

[Pi18] Pinto, Allan; Pedrini, Helio; Krumdick, Michael; Becker, Benedict; Czajka, Adam; Bowyer, Kevin W; Rocha, Anderson: Counteracting presentation attacks in face, fingerprint, and iris recognition. Deep Learning in Biometrics, 245, 2018.

[RSR15] Rattani, A.; Scheirer, W.J.; Ross, A.: Open Set Fingerprint Spoof Detection Across Novel Fabrication Materials. IEEE TIFS, 10(11):2447–2460, Nov 2015.

[SC15] Sequeira, Ana F.; Cardoso, Jaime S.: Fingerprint liveness detection in the presence of capable intruders. Sensors, 15:14615–14638, 2015.

[Se16] Sequeira, A. F.; Thavalengal, S.; Ferryman, J.; Corcoran, P.; Cardoso, J. S.: A realistic evaluation of iris presentation attack detection. In: 39th TSP. pp. 660–664, June 2016.

[To18] Tolosana, Ruben; Gomez-Barrero, Marta; Kolberg, Jascha; Morales, Aythami; Busch, Christoph; Ortega-Garcia, Javier: Towards fingerprint PAD based on cnn and short wave infrared imaging. In: 2018 BIOSIG. IEEE, pp. 1–5, 2018.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 191

A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling

Ali Khodabakhsh[1], Christoph Busch[2]

Abstract: Photo- and video-realistic generation techniques have become a reality following the advent of deep neural networks. Consequently, there are immense concerns regarding the difficulty in differentiating what content is real from what is synthetic. An example of video-realistic generation techniques is the infamous Deepfakes, which exploit the main modality by which humans identify each other. Deepfakes are a category of synthetic face generation methods and are commonly based on generative adversarial networks. In this article, we propose a novel two-step synthetic face image detection method in which general-purpose features are extracted in a first step, trivializing the task of detecting synthetic images. The anomaly detector predicts the conditional probabilities for observing every individual pixel in the image and is trained on pristine data only. The extracted anomaly features demonstrate true generalization capacity across widely different unknown synthesis methods while showing a minimal loss in performance with regard to the detection of known synthetic samples.

Keywords: Deepfake, Video Forensics, Generative Adversarial Networks, PixelCNN, Universal Background Model.

1 Introduction

Advancements in the computational capacity of modern graphical processing units (GPUs) in the past decades allowed the realization of deep neural network models. Deep learning, among other contributions, provided solutions for the synthesis of photo- and video-realistic content, challenging the existing manipulation detection methods in video forensics. An especial case of such synthetic signals is "Deepfakes", which are typically generated by generative adversarial networks (GANs). Deepfakes in combination with obfuscation in various forms have shown to be effective at fooling human subjects [Ro19].

The research community has responded to this threat by developing various detection methods. Yu et al. in [YDF18] made use of unique GAN fingerprints for the detection of fake images generated by these models. RNNs have been used for temporal-aware detection of Deepfakes by Guera et al. in [GD18]. The spectrum domain is used by Zhang et al. [ZKC19] for the detection of GAN generated images.

Most of the existing detection methods are, however, complex and have narrow applicability as they are trained to detect specific types of synthetic signals and fail to generalize [Kh18]. Few publications try to address the detection of synthetic samples from unknown generation models. In [St19], Stehouwer et al. used attention mechanisms and achieved remarkable performance over various generation techniques. Nataraj et al. [Na19] used pixel co-occurrence matrices for generalized detection across different GAN architectures. In [Ma19], Marra et al. utilized multi-task learning incrementally for detecting

[1] NTNU, IIK, Norwegian Biometrics Lab, Gjovik, NO, ali.khodabakhsh@ntnu.no
[2] NTNU, IIK, Norwegian Biometrics Lab, Gjovik, NO, christoph.busch@ntnu.no

synthetic images coming from unknown GAN models. Zhou et al. [Zh17] proposed a two-stream classification network architecture based on steganalysis features. Afchar et al. [Af18] utilized mesoscopic features along with shallow networks gaining robustness against unknown synthetic images. Rossler et al. [Ro19] evaluated different detection systems on a large dataset of diverse synthetic samples and achieved the best performance with a pretrained XceptionNet neural network. For an extensive review on the related literature, please refer to [Ve20].

Despite major progress in the detection of synthetic face images, the generalization problem across widely different generation techniques remains a major issue. In this article, we propose a novel general-purpose feature. The subsequent trivialization enables a simple detector to reliably detect unknown attacks form widely different generation techniques. The proposed method achieves this by suppressing the content of the input signal while faithfully conserving the detection-relevant information. The rest of this article is organized as follows: Section 2 explains the proposed two-step method along with the rationale behind it. Section 3 explains the experimental setup used for showcasing the performance of the method, and Section 4 discusses the findings of the article. Finally, Section 5 concludes the article.

2 Methodology

Synthetic images contain artefacts that can be used for detection and can act like fingerprints for identification of their generation process. These traces, however, are often minuscule and can be severely obscured by the actual content of the images to the extent of becoming imperceptible to the eyes of the viewer as well as the automated detection systems. We hypothesize that in the synthetic face detection task, the actual content of images acts as a strong noise, and removing them would unveil these traces and greatly simplify the task of synthetic face detection. However, this approach requires knowledge of the actual content of the image for reference.

In the absence of a reference to be subtracted from the image, the likelihood of the image to an accurate probability distribution of pristine face images would serve as a suitable proxy. To make the accurate modeling of the probability distribution over the face image space practical, the image can be broken down into smaller segments, and the probability distribution over individual segments of the image conditioned on the previous segments can be modeled.

2.1 Pixel RNN

The probability distribution of intensity values in each pixel conditioned on pixels before (in raster order) in pristine images can be modeled with a PixelRNN model [VDOKK16]. In this model, for each pixel i, the probability distribution (in the form of a Logistic mixture model) of observing the current value given all previous pixel values is learned by a recurrent or a masked convolutional neural network. This network would then be able to predict the probability distribution of pixel values for each pixel location conditioned on the pixel values before it. This probability distribution can then be used to measure the likelihood of observing a specific pixel value in location x_i given all pixel

values before it ($log(p(x_i|x_{<i}))$). By repeating this operation over all the pixels in an input image, one can calculate a likelihood matrix with the same size as the input image. Consequently, the probability of observing the input image can be calculated as $log(p(x)) = \sum_{i=0}^{n} log(p(x_i|x_{<i}))$. For the purpose of this study, an improved variant of PixelRNN named PixelCNN++ [Sa17] is used.

2.2 Classification

The probability of the input image is a feature that can spot anomalies and can directly be used for classification. However, the conditional probability matrix corresponding to the log-likelihood of observing every single pixel intensity can serve as a better feature for classification as it contains additional information with respect to the location of anomalies and the anomaly strength at each location. For achieving a higher detection rate, one can use the model trained in the previous step as an anomaly feature extractor, or in more precise terms a universal background model (UBM). The term UBM signifies that the model is universally used regardless of the synthetic method in question in the detection task. Furthermore, it signifies that the model is a background preprocessing step which postpones the classification task to a second step. Consequently, a classifier can be trained on the output of the UBM model which is in the form of a conditional probability matrix in a supervised manner. Ideally, as the complexity of the detection problem is substantially reduced following the feature extraction step, a simple classifier should be sufficient for detection of synthetic faces. In this study, we use a very simple and small neural network for classification.

2.3 Generalization Performance

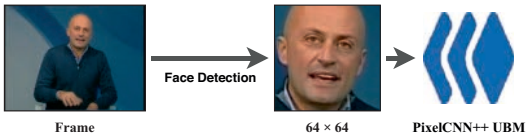
To measure the generalization capacity of a model, a common practice is to split the generation techniques to known and unknown methods. Next, the model is trained on synthetic data from the known methods and tested on the data from the unknown methods. To show the generalization capacity of our proposed method, we follow the same convention and do generalization tests in a leave-one-out (LOO) manner. For each generation method, we consider all other methods to be known and measure the detection performance on the single unknown method. The overall generalization performance is then measured by aggregating them over all the leave-one-out runs.

3 Experiment Setup

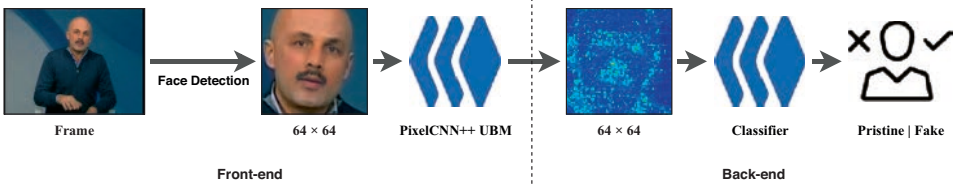
For the purpose of this study, the FaceForensics dataset [Ro19] is selected as a large dataset containing four manipulation techniques, namely Deepfakes[3], Face2Face [Th16], Faceswap[4], and Neural Textures [TZN19]. This dataset contains 1000 pristine videos along with 1000 from each manipulation technique, each split into three sets of training (with 700 videos), development (with 150 videos), and test (with 150 videos). The videos are collected from YouTube and have a minimum quality of 480p (VGA). The videos are provided in three different quality levels to simulate the conditions of video processing in

[3] https://github.com/deepfakes/faceswap
[4] https://github.com/MarekKowalski/FaceSwap/

social networks. For extraction of face images from the videos, the Dlib toolkit [Ki09] is used, and the detected face images are resized to 64×64. As the focus of this study is generalizability across the four completely different generation techniques, we limit the experiments to uncompressed data. Subsequently, the models are trained on individual cropped face images from frames as shown in Figure 1, and the detection performance is evaluated in terms of the frame-level detection accuracy.



(a) The pipeline for the training of the anomaly detection system. The model is trained on pristine face images only.



(b) The training and evaluation pipeline of the classifier. The pre-trained anomaly detection model is used as an anomaly feature extractor.

Fig. 1: The training and evaluation pipelines of the proposed method. UBM stands for universal background model and represents the probability distribution based anomaly extraction system.

The UBM model used for experiments is the Tensorflow implementation of PixelCNN++ [Sa17]. The default architecture, consisting of three blocks with five ResNet layers and 160 filters in each layer is used. A single model with 94 million parameters is trained for five epochs on natural images only from the training set, with a learning rate of 0.0001 on a single GPU in an end-to-end manner.

As the complexity of the detection problem is reduced in the anomaly feature extraction step to an extent that the synthesis artifacts are visible in its output (see Figure 4), a very simple classifier based on LeNet-5 [Le98] is used for detection of synthetic faces from known and unknown generation methods. The modified architecture summarized in Figure 2 is small enough to be trained on a CPU and has less than one million parameters. For each experiment, one classifier is trained on the available training data for 25 epochs with a learning rate of 0.001. The activation function used is the ReLU function, and to improve the convergence speed, batch normalization is used between the output of the layers and the activation function. The overall detection pipeline is shown in Figure 1b.

4 Results and Discussion

In this section, we first discuss the characteristics of the anomaly extraction method and then summarize the performance of the method on both known and unknown attack detection scenarios.

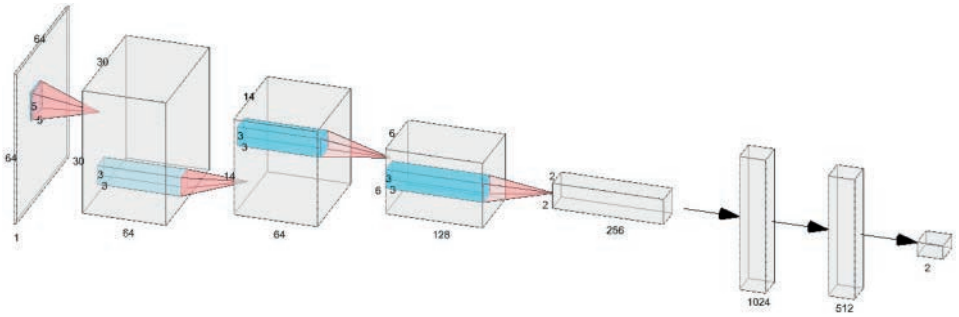


Fig. 2: The diagram of the classifier architecture. Each convolution is followed by a 2x2 maxpooling layer and ReLU activation. The network has a total of 933,442 parameters.

4.1 Features

Figure 3 shows the histogram of log-likelihoods for images in the validation data for pristine images as well as the synthetic images. The log-likelihood values for the pristine images are higher than the synthetic images, however, there is a significant overlap between the distributions. Deepfakes show higher log-likelihood values compared to the other synthesis methods. These results show the discrimination power of the observation probability of the images for synthetic face image detection. However, the image probability distributions have significant overlap, and cannot be relied on as a high-performance detection score.

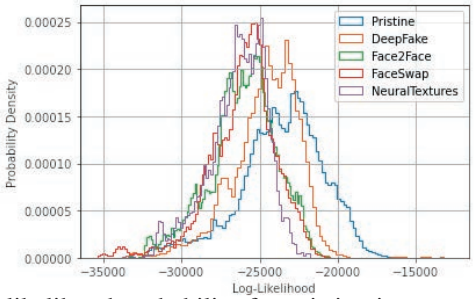
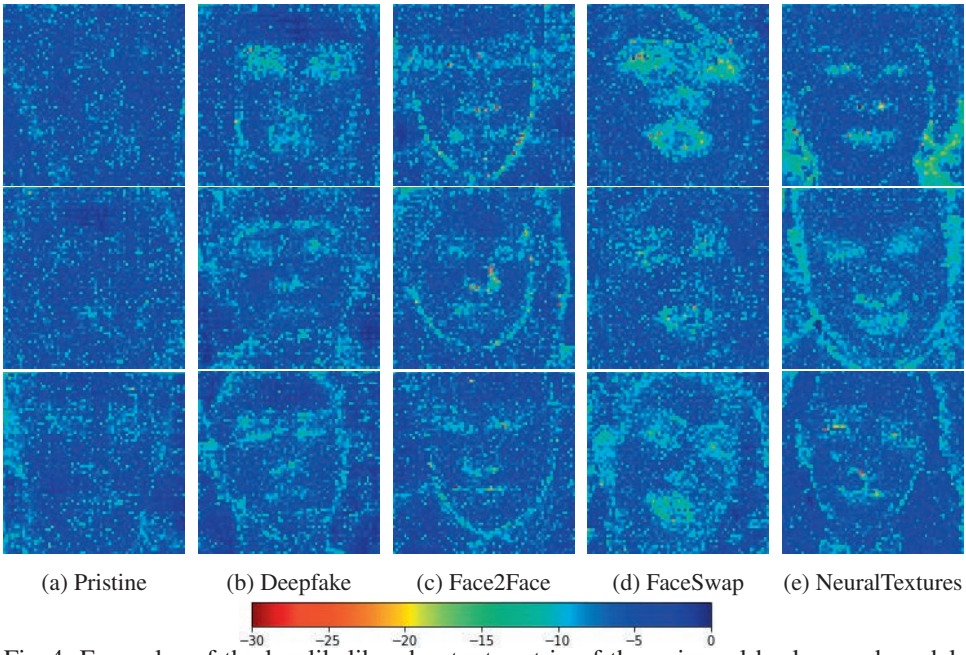


Fig. 3: The image log-likelihood probability for pristine images and synthetic images in the development data.

To achieve a better performance, we can rely on the pixel log-likelihood *images* extracted by the UBM model as anomaly features. Figure 4 visualizes examples of these *images* from the pristine data as well as the four generation techniques. In this figure, a drastic difference is observable between the pristine images and the synthetic images. The traces of the synthesis process are visible as low likelihood points in yellow and red on the image. Furthermore, each generation method shows a unique footprint in all examples. The Deepfakes have artifacts in the shape of the spliced synthetic face area over the background image. The Face2Face technique results in low likelihood pixel values on the edges of the 3D facial features such as nose and jawline. FaceSwap technique results in low likelihood areas around the eyes and the mouth. Lastly, NeuralTextures inhibits individual low-likelihood pixels on the nose and eye regions.



(a) Pristine (b) Deepfake (c) Face2Face (d) FaceSwap (e) NeuralTextures

Fig. 4: Examples of the log-likelihood output matrix of the universal background model on pristine and synthetic face images. The name of the generation method is mentioned below each column. As shown in the color bar, red signifies low log-likelihood probability, while blue signifies high.

4.2 Known Synthetic Face Detection

To measure the discriminative power of the likelihood images, we used the simple classifier explained in the previous section for synthetic face detection on each individual method. The results are reported in Table 1 along with the performance of the baseline methods from [Ro19]. The proposed method performs on par with the baseline methods despite having a smaller input image size and a much smaller number of parameters. These results confirm that the log-likelihood images conserve the information valuable for detection faithfully while reducing the detection complexity by removing the unhelpful information.

4.3 Unknown Synthetic Face Detection

The performance of the proposed method in the unknown synthetic face detection scenario is summarized in Table 2. The proposed method shows an acceptable detection rates for all four synthesis methods while showing above 96% on three out of four in LOO generalization experiments. The performance of Face2Face method gets slight improvement over the known case due to the larger training data available in the LOO scenario.

5 Conclusion

In this article, we introduced a truly generalizable synthetic face image detection method which achieves an outstanding average detection accuracy of 95.73% on unknown synthetic methods. The synthetic methods are from widely different synthesis mechanisms

Tab. 1: The performance of the proposed method in terms of detection accuracy in known synthetic face image detection scenario in comparison with existing methods adapted from [Ro19]. (DF: DeepFakes, F2F: Face2Face, FS:FaceSwap, NT:NeuralTextures)

	Input Size	DF [%]	F2F [%]	FS [%]	NT [%]
Steg. Features+SVM [FK12]	128×128	99.03	99.13	98.27	99.88
Cozzolinoet al. [CPV17]	128×128	98.83	98.56	98.89	99.88
Bayar and Stamm [BS16]	128×128	99.28	98.79	98.98	98.78
Rahmouniet al. [Ra17]	100×100	98.03	98.96	98.94	96.06
MesoNet [Af18]	256×256	98.41	97.96	96.07	97.05
XceptionNet [Ch17]	299×299	99.59	99.61	99.14	99.36
Proposed Method	64×64	99.30	98.25	99.11	98.46

Tab. 2: The performance of the proposed method on unknown synthetic samples in terms of detection accuracy. For each method, the system is trained on the other three synthesis data and did not observe a single sample of the method in question during training. The average detection accuracy is also reported. (DF: DeepFakes, F2F: Face2Face, FS:FaceSwap, NT:NeuralTextures)

	DF [%]	F2F [%]	FS [%]	NT [%]	Avg [%]
LOO Detection Accuracy	89.26	98.41	96.80	98.44	95.73

ranging from Deepfakes from generative adversarial networks to FaceSwap. The proposed method consists of a preprocessing step where the content of the image is suppressed, and the anomaly locations and anomaly strengths are extracted. The classification is then done by a simple classifier. The anomaly extraction step is trained on natural images only and preserves the detection-relevant information faithfully in the form of observation log-likelihood probability. The detectors' success provides new hopes for addressing the generalization problem over widely different generation processes.

References

[Af18] Afchar, Darius; Nozick, Vincent; Yamagishi, Junichi; Echizen, Isao: Mesonet: a compact facial video forgery detection network. In: WIFS. IEEE, pp. 1–7, 2018.

[BS16] Bayar, Belhassen; Stamm, Matthew C: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: ACM IH&MMSec. pp. 5–10, 2016.

[Ch17] Chollet, François: Xception: Deep learning with depthwise separable convolutions. In: IEEE CVPR. pp. 1251–1258, 2017.

[CPV17] Cozzolino, Davide; Poggi, Giovanni; Verdoliva, Luisa: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: ACM IH&MMSec. pp. 159–164, 2017.

[FK12] Fridrich, Jessica; Kodovsky, Jan: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 7(3):868–882, 2012.

[GD18] Güera, David; Delp, Edward J: Deepfake video detection using recurrent neural networks. In: IEEE AVSS. IEEE, pp. 1–6, 2018.

[Kh18] Khodabakhsh, A.; Ramachandra, R.; Raja, K.; Wasnik, P.; Busch, C.: Fake Face Detection Methods: Can They Be Generalized? In: BIOSIG. pp. 1–6, 2018.

[Ki09] King, Davis E: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10(Jul):1755–1758, 2009.

[Le98] LeCun, Yann; Bottou, Léon; Bengio, Yoshua; Haffner, Patrick: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[Ma19] Marra, Francesco; Saltori, Cristiano; Boato, Giulia; Verdoliva, Luisa: Incremental learning for the detection and classification of GAN-generated images. arXiv preprint arXiv:1910.01568, 2019.

[Na19] Nataraj, Lakshmanan; Mohammed, Tajuddin Manhar; Manjunath, BS; Chandrasekaran, Shivkumar; Flenner, Arjuna; Bappy, Jawadul H; Roy-Chowdhury, Amit K: Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging, 2019(5):532–1, 2019.

[Ra17] Rahmouni, Nicolas; Nozick, Vincent; Yamagishi, Junichi; Echizen, Isao: Distinguishing computer graphics from natural images using convolution neural networks. In: WIFS. IEEE, pp. 1–6, 2017.

[Ro19] Rossler, Andreas; Cozzolino, Davide; Verdoliva, Luisa; Riess, Christian; Thies, Justus; Nießner, Matthias: Faceforensics++: Learning to detect manipulated facial images. In: IEEE ICCV. pp. 1–11, 2019.

[Sa17] Salimans, Tim; Karpathy, Andrej; Chen, Xi; Kingma, Diederik P: Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. arXiv preprint arXiv:1701.05517, 2017.

[St19] Stehouwer, Joel; Dang, Hao; Liu, Feng; Liu, Xiaoming; Jain, Anil: On the detection of digital face manipulation. arXiv preprint arXiv:1910.01717, 2019.

[Th16] Thies, Justus; Zollhofer, Michael; Stamminger, Marc; Theobalt, Christian; Nießner, Matthias: Face2face: Real-time face capture and reenactment of rgb videos. In: IEEE CVPR. pp. 2387–2395, 2016.

[TZN19] Thies, Justus; Zollhöfer, Michael; Nießner, Matthias: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG), 38(4):1–12, 2019.

[VDOKK16] Van Den Oord, Aäron; Kalchbrenner, Nal; Kavukcuoglu, Koray: Pixel Recurrent Neural Networks. In: ICML - Volume 48. JMLR.org, p. 1747–1756, 2016.

[Ve20] Verdoliva, Luisa: Media Forensics and DeepFakes: an overview. arXiv preprint arXiv:2001.06564, 2020.

[YDF18] Yu, Ning; Davis, Larry P; Fritz, Mario: Attributing fake images to gans: Analyzing fingerprints in generated images. 2018.

[Zh17] Zhou, Peng; Han, Xintong; Morariu, Vlad I; Davis, Larry S: Two-stream neural networks for tampered face detection. In: CVPRW. IEEE, pp. 1831–1839, 2017.

[ZKC19] Zhang, Xu; Karaman, Svebor; Chang, Shih-Fu: Detecting and simulating artifacts in gan fake images. arXiv preprint arXiv:1907.06515, 2019.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 199

Eyebrow Recognition for Identifying Deepfake Videos

Hoang (Mark) Nguyen[1], Reza Derakhshani[2]

Abstract: Deepfake imagery that contains altered faces has become a threat to online content. Current anti-deepfake approaches usually do so by detecting image anomalies, such as visible artifacts or inconsistencies. However, with deepfake advances, these visual artifacts are becoming harder to detect. In this paper, we show that one can use biometric eyebrow matching as a tool to detect manipulated faces. Our method could provide an 0.88 AUC and 20.7% EER for deepfake detection when applied to the highest quality deepfake dataset, Celeb-DF.

Keywords: Deepfake detection, eyebrow biometrics, biometric recognition.

1 Introduction

In recent years, digital media is playing an exceedingly influential role in different aspects of our lives, including shaping public opinion. More and more people are getting their information from social networks and video-sharing platforms. Unfortunately, technology has also allowed images and videos to be manipulated by nefarious actors to show misinformation and discord. This issue has become a public concern threatening information trustworthiness and even undermining democracies [Ci19]. The tools for manipulating imagery, such as those used for political misinformation, have become widely available [VC20, Ag19].

The now-famous term ”deepfake” refers to recent (deep-learning-based) techniques used to synthesize or otherwise alter imagery, mostly faces in videos, which is also the focus of this paper. Due to rapid advances in computer vision and with increasingly affordable and capable hardware, convincing fake visual contents are being created and distributed at an alarming rate. Recently we have seen deepfake videos seeding misinformation by depicting public figures uttering words they had never said, among many other egregious and vulgar applications. As a result, deepfake detection is quickly becoming a high priority topic for the research community, the industry, and the governments alike.

Current anti-deepfake algorithms heavily rely on detecting image or video abnormalities such as visible artifacts or lack of coordination between lip movements and spoken words. Some examples of the aforesaid facial artifact are shown in Fig 1. During facial synthesis, many deepfake generators extract facial landmarks from the videos to manipulate the facial areas of interest. After manipulating the targeted facial features, a series of post-processing

[1] Department of Computer Science and Electrical Engineering, University of Missouri at Kansas City, hdnf39@mail.umk.edu
[2] Department of Computer Science and Electrical Engineering, University of Missouri at Kansas City, derakhshanir@umkc.edu

methods such as resolution-enhancement and color correction are applied to render the manipulated visualizations more realistic. Facial manipulation methods may be applied to the entire face or just the parts needed for the facial expressions [To20]. However, as the deepfake technologies improve over time, the deepfake visualization has become more realistic, and fewer artifacts are visible in the altered images and videos. REFACE app[3] is a face swap mobile application and has successfully integrated user face into high-quality music videos. Celeb-DF [Li20] is the highest quality deepfake forensic dataset that is publicly available. Fig 2 shows examples from the dataset. The current state-of-the-art deepfake detection approaches have not performed very well on this dataset, given its high quality (table 1). Thus, instead of relying on the hard to find visible artifacts for such datasets, we utilize a biometric recognition model to distinguish between real and fake images from the same identity.

In this paper, our focus is on detecting face swap by matching the components of the swapped face. More specifically, we show the efficacy of matching the eyebrow area to counter deepfake attacks. One may add other components like lower periocular to such a system. As the deepfake algorithms improve over time, one can expect the altered image artifacts to vanish, and thus biometric comparison of the swapped components may be preferable to flag counterfeit imagery.

The main contribution of this work is establishing the utility of eyebrows for deepfake detection by way of biometric comparison. To the best of our knowledge, this is the first time that biometric comparison of the eyebrow region is proposed for deepfake detection. The eyebrow region is one of the most affected components in the synthesized images. Especially in high-resolution and high-quality deepfakes, we show that eyebrow alterations become more distinguishable if examined by a biometric comparison pipeline. In order to make this approach to work, the model needs to know the participant's identity beforehand (biometric enrollment is needed). Moreover, this will be applicable when the targets are well-known individuals are celebrities or politicians.

Fig. 1: Examples facial artifacts in deepfake database

2 Prior Work

[MRD19] is one of the most recent works in eyebrow recognition. Mohammad et al. investigated short term eyebrow recognition using VISOB and FERET datasets. The au-

[3] https://reface.app/

Fig. 2: Examples from celeb-DF dataset: (a) real images, (b) deepfake images. Images in each column belong to the same identity

thors proposed a fusion of GIST, HOG, and VGG16 features along with Support Vector Machine (SVM) classifiers for biometric comparison. 0.63% Equal Error Rate (or EER, lower is better) and 0.9942 Area Under the Curve (or AUC, higher is better) was their best-reported results when fusing three feature descriptors for both eyebrows. However, their evaluates followed a closed set protocol where there are overlaps between training and testing set identities.

[MRS19] exploit visual artifacts in images to detect deepfakes. The authors proposed various facial areas where their model could spot potential artifacts caused by manipulating facial imagery. Some examples of such artifacts are global inconsistencies, illumination mismatches, geometrical distortions (such as those observed over the teeth), and eye color issues. Their best-reported results is 0.866 AUC using their in-house dataset.

[St19] proposed using an attention mechanism to detect manipulated face images. The attention map guides a CNN to scrutinize the face region in the image. The attention map mask helps the elimination of irrelevant features and thus reduces the feature vector dimensionality. Therefore, only certain sub-region in the vicinity of the face make significant contributions to the CNN's decision. The proposed approach reportedly achieves a 0.984 AUC in the UADFV dataset and 0.712 AUC over the Celeb-DF dataset.

Table 1 summarizes reported deepfake detection results over the Celeb-DF dataset. As mentioned earlier, to the best of our knowledge, this paper's proposed method is the first work using an eyebrow biometric pipeline to counter deepfake attacks. It is also noteworthy that we did not train our biometric model on any deepfake datasets, saving them for eventual testing to demonstrate cross-dataset generalization.

Tab. 1: performance of recent deepfake detection on Celeb-DF dataset.

Ref	Detection Method Used	Classifiers	Best AUC
Zhi et al. (2018) [Zh17]	Image-related Steganalysis	CNN+SVM	538
Afchar et al. (2018) [Af18]	Mesoscopic Level	CNN	0.548
Yang et al. (2018) [YLL19]	Head Pose Estimation	CNN	0.546
Li et al. (2019) [LL18]	Face Wrapping Artifact	CNN	0.569
Matern et al. (2019) [St19]	Visual Artifact	Losgistic Regression MLP	0.551
Stehouwer et al. (2019) [MRS19]	Facial Forgery	Attention Mapping	0.712

3 Methods

We employed four deep learning models to evaluate our hypothesis: LightCNN, Resnet, DenseNet, and SquezeNet. They are widely used in biometric research publication. Therefore, we believe that they would achieve high performance in eyebrow matching task.

LightCNN LightCNN [WHS15] model heavily relies on Max-Feature-Map (MFM) operation which was proposed in place of ReLu activation function. The operation preserves element-wise maximum from two feature maps forcing only half of the features to reach the next layer. In other words, this acts as a filter allowing the only compact feature to pass through.

ResNet Resnet [He16] employs a shortcut connections to deal with the gradient degradation problem[GB10]. Such an issue happens when training very deep neural networks. The residual or shortcut connections introduced in ResNet allows for identity mappings to propagate to multiple nonlinear layers, preconditioning the optimization during training. In this paper, we used ResNet-50 consists of 49 convolution layers and a single fully connected layer.

DenseNet The unit's dense block was first introduced in Dense Convolutional Network or DenseNet [Hu17]. In each block, there are multiple convolution layers where each layer is a concatenation of feature maps from previous layers. 1x1 convolutions are also utilized to reduce a large number of feature maps and the computation complexity. In this work, the DenseNet-121 model is utilized.

SqueezeNet Iandola1 et al. proposed SqueezeNet[Ia16], an efficient model, which is 50 times smaller than AlexNet but achieved the same level of accuracy. The model employs many strategies to decrease the number of parameters, such as small filter size, reduced input channels, and squeezed layers. Fire module was also introduced in the paper consisting of two layers: a squeeze layer consisting of 1×1 convolution filters, and expand layer, which is a mix of 1×1 and 3×3 convolution filters. The module does not decrease only the number of 3×3 filters but also the input channel.

Matching After obtaining our models' feature vectors, we used cosine distance metric to measure the similarity between reference and probe eyebrows. This is a famous match score employed by many deep-learning-based biometric systems.

4 Experimental Evaluation

Training Data VISible light mobile Ocular Biometric (VISOB) [Ra16] is a publicly available dataset consisting of eye images of about 550 healthy adults captured by three different mobile phones in three different lighting conditions. The three smartphones used in data collection are OPPO N1, iPhone 5s, and Galaxy Note 4. During the data collection, the volunteers were asked to take selfie-like images during two visits (Visit 1 and Visit 2), 2-4 weeks apart. During each visit, images were taken in two sessions 10-15 minutes apart, and under three different illumination conditions: regular office light, dim indoors, and natural daylight. In this experiment, we used a subset of VISOB captured under office lighting using the OPPO device, which offers a better resolution than iPhone and Note 4 captures, for our model training. Our model was trained on a high-resolution subset of VISOB to tell apart identities by way of eyebrow matching.

Testing Data Celeb-DF is the large, high quality deepfake forensic dataset. This dataset consists of 590 real videos from 59 celebrities along with 5639 deepfake videos. Since we are not after visual artifact caused by image synthesis, we evaluated our model on the best quality deepfake dataset that provides the most realistic fake video. This is a challenging task that nonetheless can better demonstrate the advantages of our proposed method. Unlike the other datasets, Celeb-DF contains almost no splicing boundaries, color mismatch, and inconsistencies of face orientation, among other visible deepfake artifacts. As a result, several deepfake detection papers have reported low accuracy numbers on this dataset. As shown in table 1, the current detection methods peak around 75% AUC on this dataset.

Data processing and training setup : We divided the VISOB dataset into 80% for training and 20% for validation. The eyebrow images were resized to different sizes depending on the corresponding deep learning models' input requirements. Multiple augmentations such as random rotations and random cropping, were applied to the training set. We trained our models with an initial learning rate of $1e^{-3}$ and reduced it by ten if the validation loss

did not drop by ten consecutive epochs. We trained our model for a maximum of 200 epochs and ended with the weights from the epoch that yielded the best validation loss. The momentum and weight decay parameters were set to 0.9 and $10e^{-4}$, respectively.

Experimental Setup We chose two experiments, short term and long term evaluation, to evaluate our hypothesis. All genuine matches in the former came from different frames in the same real video, while genuine matching was performed across the real videos in the latter. For each celebrity, one video out of ten videos was chosen to perform genuine matching in short-term evaluation. On the other hand, for the long term experiment, we used all the real videos for evaluation. For both experimental setups, all the deepfake videos were included to perform imposter matching with the real videos. For both the experiments, we extracted one frame from each deepfake video, and 20 frames from each real video (10 for enrollment and 10 for verification). The genuine match score is calculated between two images from the real video, and the imposter match score is calculated between a frame from the real video and another frame from deepfake video. We only perform matching between the original video and synthesized video from the same identities. These experiments are completely open-set that the participants in the training set are not from the identities used in the testing set. Further, the fake vs. real evaluations are conducted within the same quality and samples enjoy comparable resolutions regardless of their class label. We used ROC's Equal Error Rate (EER%) and Area Under the Curve (AUC) metrics to convey accuracies.

Tab. 2: EER and AUC for short term eyebrows identification in real and deep fake imagery

	Model	lightCNN	ResNet	DenseNet	SqueezeNet
Left	AUC	0.729	0.762	0.700	0.832
	EER	31.8%	29.5%	35.7%	25.3%
Right	AUC	0.696	0.879	0.690	0.802
	EER	35.4%	20.7%	37.6%	28.0%

5 Results and Discussions

Table 2 shows the EER and AUC for our short term evaluation. The best-achieved accuracy is 20.7% EER and 0.879 AUC using ResNet on the right eyebrow. For the left eyebrow, SqueezeNet performed the best with 25.0% EER and 0.832 AUC (its corresponding results for the right eyebrow were 28.0% EER and 0.802 AUC). The worst performer was DenseNet with 0.690 AUC and 37.6% EER (right eyebrow).

The accuracies for our long term evaluation are summarized in table 3. As expected, these results are worse than the short term's results with AUCs from 0.548 to 0.589 and EERs around 45.0%. This indicates that eyebrow matching is not the best choice for long term comparisons.

Tab. 3: EER and AUC for long term eyebrows identification in real and deep fake imagery

	Model	lightCNN	ResNet	DenseNet	SqueezeNet
Left	AUC	0.597	0.567	0.563	0.573
	EER	44%	45.3%	45.1%	45.3%
Right	AUC	0.589	0.580	0.548	0.561
	EER	43.3%	43.4%	46.6%	45.3%

6 Conclusion and Future Work

With the rapid developments in image synthesis, the creation of convincing deepfake videos has become easier and readily available to almost many. Since most of the deepfake detection methods rely on visible structural artifacts or color inconsistencies, they do not perform well on high-quality deepfake datasets such as Celeb-DF. In this work, we showed the efficacy of a new approach to expose deepfake images or videos using eyebrow matching. Instead of detecting the visible signs of facial manipulation, we used eyebrow match scores between real versus fake images from the same identity. Our best-achieved accuracy was 20.7% EER and 0.879 AUC on Celeb-DF, which is significantly better than other recently reported results on this high-quality deepfake dataset. However, we also noted that our approach did not fare as well over long term evaluations. Another limitation of our method is the requirement for the subject's identity so that the biometric eyebrow matching can proceed. As a part of future work, we would like to utilize the more feature-rich continuous eyebrow band region (simultaneously presenting both eyebrows) with our approach. Lastly, although our evaluations were made on a dataset different from the development set, we wish to perform additional cross-dataset deepfake evaluations to further test the generalization capability of the proposed framework.

7 Acknowledgement

This work was made possible in part by a gift from ZOLOZ. Dr. Derakhshani is also a consultant for the company.

References

[Af18] Afchar, Darius; Nozick, Vincent; Yamagishi, Junichi; Echizen, Isao: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, pp. 1–7, 2018.

[Ag19] Agarwal, Shruti; Farid, Hany; Gu, Yuming; He, Mingming; Nagano, Koki; Li, Hao: Protecting world leaders against deep fakes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 38–45, 2019.

[Ci19] Citron, Danielle: , How DeepFake Undermine Truth and Threaten Democracy, 2019.

[GB10] Glorot, Xavier; Bengio, Yoshua: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256, 2010.

[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.

[Hu17] Huang, Gao; Liu, Zhuang; Van Der Maaten, Laurens; Weinberger, Kilian Q: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708, 2017.

[Ia16] Iandola, Forrest N; Han, Song; Moskewicz, Matthew W; Ashraf, Khalid; Dally, William J; Keutzer, Kurt: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 MB model size. arXiv preprint arXiv:1602.07360, 2016.

[Li20] Li, Yuezun; Sun, Pu; Qi, Honggang; Lyu, Siwei: Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In: IEEE Conference on Computer Vision and Patten Recognition (CVPR). Seattle, WA, United States, 2020.

[LL18] Li, Yuezun; Lyu, Siwei: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656, 2018.

[MRD19] Mohammad, A. S.; Rattani, A.; Derakhshani, R.: Eyebrows and eyeglasses as soft biometrics using deep learning. IET Biometrics, 8(6):378–390, 2019.

[MRS19] Matern, Falko; Riess, Christian; Stamminger, Marc: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, pp. 83–92, 2019.

[Ra16] Rattani, A.; Derakhshani, R.; Saripalle, S. K.; Gottemukkula, V.: ICIP 2016 competition on mobile ocular biometric recognition. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 320–324, Sept 2016.

[St19] Stehouwer, Joel; Dang, Hao; Liu, Feng; Liu, Xiaoming; Jain, Anil: On the detection of digital face manipulation. arXiv preprint arXiv:1910.01717, 2019.

[To20] Tolosana, Ruben; Vera-Rodriguez, Ruben; Fierrez, Julian; Morales, Aythami; Ortega-Garcia, Javier: DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. arXiv preprint arXiv:2001.00179, 2020.

[VC20] Vaccari, Cristian; Chadwick, Andrew: Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media+ Society, 6(1):2056305120903408, 2020.

[WHS15] Wu, Xiang; He, Ran; Sun, Zhenan: A Lightened CNN for Deep Face Representation. CoRR, abs/1511.02683, 2015.

[YLL19] Yang, Xin; Li, Yuezun; Lyu, Siwei: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 8261–8265, 2019.

[Zh17] Zhou, Peng; Han, Xintong; Morariu, Vlad I; Davis, Larry S: Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 1831–1839, 2017.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 207

Face Presentation Attack Detection in Ultraviolet Spectrum via Local and Global Features

Dirk Siegmund, Florian Kerckhoff, Javier Yeste Magdaleno, Nils Jansen,
Florian Kirchbuchner[1], Arjan Kuijper[2]

Abstract: The security of the commonly used face recognition algorithms is often doubted, as they appear vulnerable to so-called presentation attacks. While there are a number of detection methods that are using different light spectra to detect these attacks this is the first work to explore skin properties using the ultraviolet spectrum. Our multi-sensor approach consists of learning features that appear in the comparison of two images, one in the visible and one in the ultraviolet spectrum. We use brightness and keypoints as features for training, experimenting with different learning strategies. We present the results of our evaluation on our novel Face UV PAD database. The results of our method are evaluated in an leave-one-out comparison, where we achieved an APCER/BPCER of 0%/0.2%. The results obtained indicate that UV images in presentation attack detection include useful information that are not easy to overcome.

Keywords: Face Presentation Attack Detection PAD, Ultraviolet, MFP, Biometrics.

1 Introduction

Face recognition (FR) is the most commonly used biometric method for recognizing people. Applications range from unlocking smartphones and border-control to dynamic recognition in surveillance scenarios. The accuracy of face verification systems has improved significantly since the advent of deep learning, especially in scenarios where sample and probe image are taken in similar conditions. While the accuracy of FR improved, their vulnerability to presentation attacks remains a major challenge. Presentation attacks are defined as "presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system."[In16]. They range from very simple low effort attacks like printed face images or replayed videos to more sophisticated attacks involving high quality disguises and masks. Presentation attack detection (PAD) are approaches to prevent presentation attacks from single or series of images, using different properties like: motion, texture or life signs. There is currently no detection method that is absolutely safe. Especially three-dimensional masks and high quality 3D-prints can very often overcome PAD. Commonly known methods include additional analysis of images captured in different wavelengths, especially in the infrared (IR) and near-infrared (NIR). Their general vulnerability in practive is that 2D and 3D face images of people are commonly available, or can be generated even from a single image [ASJT17]. In this paper

[1] Fraunhofer Institute for Computer Graphics Research (IGD), Fraunhoferstrasse 5, 64283 Darmstadt, Germany {dirk.siegmund, florian.kerckhoff, javier.yeste.magdaleno, nils.jansen, florian.kirchbuchner}@igd.fraunhofer.de

[2] Technische Universität Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany, arjan.kuijper@tu-darmstadt.de

we present a solution that tackles PAD, by using multi-modal biometrics in the ultra-violet (UV) spectrum by analyzing Melanin Face Pigmentation (MFP). As discovered in a recent paper [Sa18], MFP can be seen as additional modality, of which most can only been seen by a sensor, sensible in UV wavelength. In this paper we analyze if these captures of the human skin are useful for PAD by presenting novel methodology. Our first method detects MFP in the images using ORB keypoints and identifies attacks using their number and distribution. In the second method, we examine whether the corresponding brightness can be used as a feature between the images. Both methods use two captures at the same time, one in the UV spectrum and another made in the visual spectrum (VIS). To confirm our assumption that PAD works by using images in the UV spectrum, we present a database of presentation attacks that includes images in UV and VIS spectrum (see Section 3). This database contains images of 2D prints on paper, 3D prints and masks of different material. These images are evaluated together with a recently published database of 91 real subjects captured over a period of 6 months showing different expressions and poses. The methodology of our verification methods is presented in Section 4. There, we describe the image descriptors and fusion methodology that we used in our methodology. Our results in Section 5 show if UV face imaging and/or MFP, provide valuable distinct information for face PAD. We conclude with a future perspective about the use of these properties for future research and highlight observed issues and limitations in Section 6.

2 Related Work

Active imposter presentation attack detection algorithms can be categorized into hardware and software based. Software based algorithms are cheaper, space saving and include static and dynamic algorithms. They can analyze micro-textural patterns [RB17] and/or motion [De12] but mostly fail when a trained model is used in a different environment or on other datasets. Damer et al. [DD16] reported good results in a motion magnification based approach using histograms of oriented optical flow. A limitations of this approach is the human physiological rhythm itself and computational costs. Hardware-based multi-spectral algorithms analyze several images in distinct regions of the electro-magnetic spectrum individually [Ra17]. There are also multi-sensor approaches, where multiple spectral bands are being used at the same time by different sensors. The spectral band can be divided into the VIS [400nm - 700nm], IR [780nm - 15 μm], NIR and the short-wave (SWIR) band. Multi-sensor/cross model approaches can take advantage of the different reflection properties of material in different spectra. In other words, knowing that human skin reflects IR light quite different than e.g. silicon, enabled the detection presentation attacks by a comparison of images which capture both spectra. The effectiveness of this method is demonstrated by the known FaceID, used in the Apple iPhone's. But while active IR or NIR images show advantages especially in robustness to illumination and exhibit special characteristics of the human skin, they can be spoofed as well by using a 3D mask[SKJ16]. Due to the MFP ascribed properties, we think that these features should also be useful for PAD.

3 Database

The evaluation of the proposed method is carried out on an extended version of a newly created UV-Face database [Sa18]. The database consists of images collected in the UV, as well as in the VIS spectrum under conditions, as one would expect them in a controlled scenario, such as border control. Compared to the IR bandwidth, one of the first observations when exposing human skin to UV emission is, that skin of different people looks quite differently in that spectra. The Fitzpartick scale[Fi88] groups the skin type into six different categories, according to the reaction of the skin to the sun. Most notable with skin type I, where people show additional MFP in the UV image, that aren't visible for human eyes. We captured 476 images of 28 identities of Skintype I and II. 1042 images of 45 identities of skintype III and IV and 330 images of 18 identities of skintype V and VI. The database includes subjects of different age, gender and skin types. We've expanded the database by 127 images of spoofing attacks by using a variety of materials based on a selection of attacks according to reported attacks in media and research. We used eight different types of masks (painted and unpainted latex and latex foam), bursts (silicone, photopolymer and PLA) and paper printouts on different paper. Each attack is captured by using both cameras

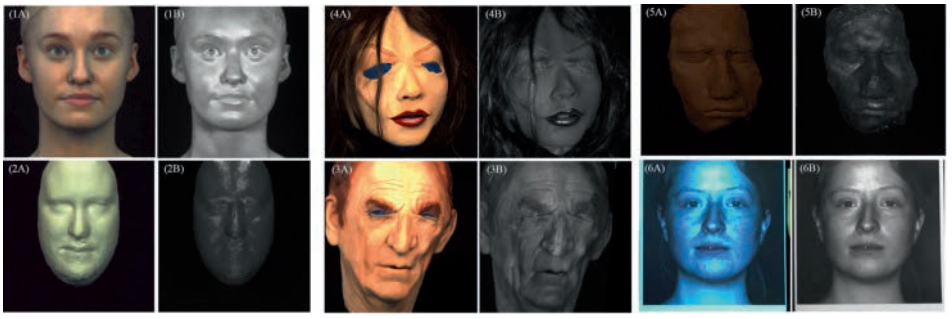


Fig. 1: Created spoofing attacks VIS (A) UV (B). (1) Color-bust made of photopolymers, by Stratasys - Connex 3 3D printer (Polyjet) (2) 3D face bust (17x11cm), by Prusa i3 MK3 3D (Polylactide) (3/4) Unpainted, professional latex masks, two painted masks of the same material and variations with wigs (5) 3D face bust (silicone rubber) on a 3D mold, using alginat for the imprint (6) Twenty laser color-printouts, Ten using normal paper, ten on thicker shiny

in following poses: frontal, 45°view to the left, 45°view to the right, looking up, looking down. Two cameras, attached side by side, are used in order to keep the divergence in perspective small. Test participants, wearing the masks, or the 3D models are positioned at a distance of 1.5m away from the cameras. In order to avoid interferences, UV/IR and VIS filters were used respectively, allowing only the transmission of the intended wavelength. For the UV capturing, a DLP LLC camera with a CMOS sensor, resulting in images of 2592x1944 pixel resolution is used. For illumination we used two 36W UV-A LPS lamps with a bandwidth between 315nm and 400nm positioned in front left and front right to the subject. The position of the used lights was chosen in a way that shades are similar in both captures. The images in the visible spectrum are captured by using a Nikon D9000 with a APS-C CMOS sensor and a 35mm lens. The UV images are resized by 58% and cropped to 600x600pixel, VIS images respectively. All images are converted to gray-scale.

We augmented the attack database by slightly changing the saturation for every image pair by using linear transformation.

4 Introduced Methods for UV-PAD

As one of the first observations, after capturing the attack images, we found that the brightness of the images differs greatly in UV compared to the VIS. Since both VIS and UV image are taken simultaneously by us, we can rule brightness manipulation by the attacker out. While the brightness of the silicone bust (see Figure 1-5B) is relatively low, the 3D color print made of photopolymers 1-1B) reflects a lot and is therefore very bright. Of course, it can also be assumed that UV images of non-skin have no MFP, which would be additionally evaluable on the UV images. Another observation is that relatively smooth material, such as latex masks with no notches, have almost no details in the UV spectrum (see Figure 1-4B). Furthermore, all latex masks show no reflections that lead to overexposure at all. Comparing that to bona fide images we observed that there is almost no image that does not show at least a small area like this (very often at the forehead). However, smooth material such as the silicon print, the 2D prints or the PCL 3D print have very strong reflections of this kind. In the case of the 2D printouts, it was even only possible at certain angles to capture images at all, where not the complete face is superimposed by this effect. The main difference between two images is the overexposure in some places, apparently due to the material. However, this effect also occurs in the images of the bona fide group, and is therefore not suitable for a targeted evaluation. These observations lead

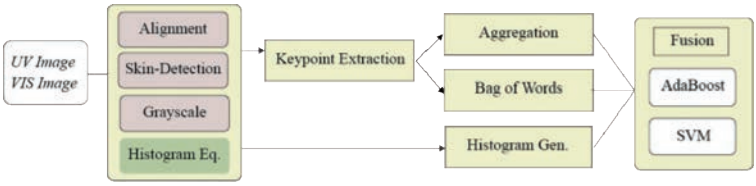


Fig. 2: Flow-Diagram of the proposed Methodology.

us to evaluate these properties in two different ways. If there are no differences between the two images, as would be the case with real skin (MFP), this is an important feature, which can be characteristic of attacks and bona fide. Secondly, there are differences in the ratio of the brightness from VIS to UV, which may differ from those of the skin, they can be seen as spectral signatures. As our database is relatively small, we could not effectively use any deep- or transfer learning approaches. Therefore, we chose conventional features to analyze those characteristics and prove their significance. Since both properties only affect the skin, we use the same preprocessing steps for both methods , which is described in the next section. Our method for extracting the different details of both images are explained in Section 4.2. The brightness differences are presented in Section 4.3.

4.1 Pre-processing

Initially, face detection is performed on the full resolution images. After that, VIS and UV images are aligned to the face region by using face alignment by Zhang et al.[Zh16]. We aligned several images manually in order to guarantee their meaningful inclusion into the dataset. Since it is not expected that the eye region will provide valuable information, we remove this region with a mask. Since hair and the mouth region also contain no valuable information, we perform skin detection by using the procedure of Buza et al. [BAO17] and mask-out all non-skin pixel. In a next step, we convert all images to grayscale, in order to reduce the complexity of our small data-set. In our approach, which evaluates the similarity of local features (See Section 4.2), we also do histogram equalization, which we do not do in the case of brightness analysis (See Section 4.3).

4.2 Analysis of Similarity using local Features

As already shown in previous work[Sa18], MFP features can be extracted effectively via keypoints (KP). We expected to find these properties which are visible in the UV spectrum in high frequency features with a pixel size between 3 and 20 pixels (px). We have therefore selected the ORB (Oriented FAST and Rotated BRIEF) feature detector [Ru11] to extract this property. The ORB detector is computationally very efficient with similar matching performance to SIFT but less affected by image noise and can be used in real-time. A maximum of 1000 ORB KP are calculated using the harris score ranking and four points to produce the oriented BRIEF descriptor. Matching is done by using the euclidean distance between two points, one in the UV image, one in the VIS image, assuming that they denote the same feature if the euclidean distance is smaller than 10 px. In Figure 3 the results of the KP extraction and matching is shown on two images. In the upper images of an attack, with unpainted latex mask, it can be seen that hardly any of them are detected on the surface. It can only be found along the mouth, while in the bona fide image (below) they are recognized throughout all many of them can be matched. The overexposed area on the forehead in the UV image is also clearly visible. With the described method we

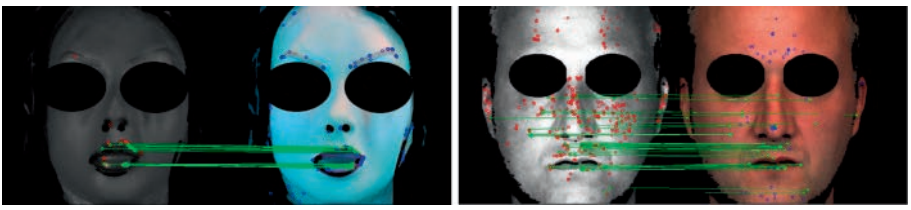


Fig. 3: Matched and Unmatched Keypoints in a typical bona fide image of Skintype II (bottom) and an Attack using a Silicone Mask (top).

have extracted the keypoints on all image pairs. We can detect the following three main differences between attacks and bona fide: (1) The number of detected keypoints is smaller for attacks compared to non-attacks (see Figure 3A)(2) Bona fide show more KP on the UV image that can't be matched with ones on the VIS image (see Figure 3B). (3) In attacks, more unmatched KP can be found on the VIS image than are found on the UV image. The

3D silicone imprint exhibits an extremely high number of unmatched keypoints on both images. These attributes allow us to distinguish both classes in particular, we visualized the number of unmatched KPs in the UV and the VIS image over all classes in Figure 4. We assume that the number of unmatched keypoints between UV and VIS, as well as between VIS and UV contain discriminative information. Therefore, we compose our feature vector as follows: (1) Total KP detected in UV (2) Total KP detected in VIS (3) Matched keypoints (4) Unmatched KP in the UV image and (5) the number of KP in VIS that couldn't be matched to the UV image.

4.3 Analysis of Brightness Property

As can be seen in Figure 3, attacks reflect differently from bona fide faces when captured with an UV camera. Figure 4 (Left) depicts the average difference between VIS and UV images of both bona fide faces and attacks presented in a gray-scale histogram. Thus, this method utilizes the distribution of their brightness values in the form of histograms. It aims to discern legitimate images from attacks by comparing the histograms of both the UV and the VIS image of faces. Since the histograms represent the image's brightness distribution, each has a length of 255. By combining both histograms for one face, we create feature vectors containing the amount of pixels that are of each particular brightness for both the UV and VIS image. We experiment with different methods of combining, including adding, subtracting and concatenating the histograms for feature vectors of a length of either 255 or 510.

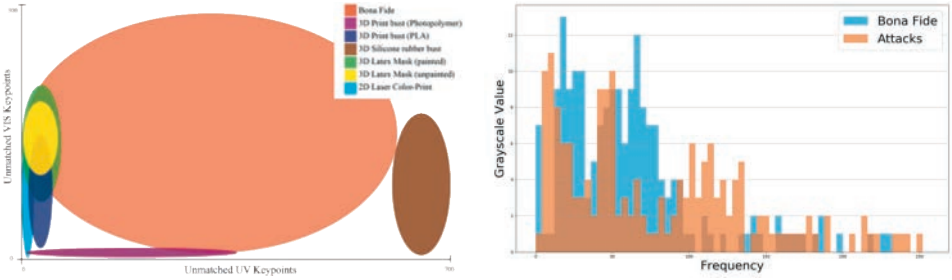


Fig. 4: (Left) Unmatched ORB Keypoints in the UV and VIS Image. (Right) Histogram comparison of Attacks(red) and Bona Fide (blue) in Grayscale.

5 Experiments and Results

In our approach using keypoints, we experimented with adding and omitting the features written in Section 4.2. Here, the variant using all five values has proved to be the best. Using the positions of the keypoints, we experimented with different feature vector lengths between 150 and 500. A length of 300 has proven to be optimal. The feature vectors of the histogram approach are created by either concatenating, adding or subtracting the histograms of UV and VIS photo for a total of 9 experiment setups. The different setups

are then evaluated based on APCER[3] and BPCER[4]. While the SVM and AdaBoost approaches both yielded usable results (with the AdaBoost approach performing the best). The logistic regression approach was not able to capture the difference of legitimate faces and attacks to an acceptable degree. This is likely due to the high amount of data required to train neural networks in comparison to SVM and AdaBoost. Among the different vector combination approaches, adding and concatenating performed comparatively (adding performing slightly better), while subtracting did not perform as well, likely due to a loss of information when brightness value resulted in zero.

Tab. 1: Our Results on the presented Dataset.

Scenario	Histogram		Keypoints		Fused	
	APCER	BPCER	APCER	BPCER	APCER	BPCER
Only Skintype 1-2	0%	0.4%	2.2%	2.45%	0%	0%
Only Skintype 3-4	0%	0.4%	3.3%	3.0%	0%	0%
Only Skintype 5-6	0%	0.4%	3.9%	6.9%	0%	0.2%
All	0.4%	1.2%	4.2%	7.2%	0%	0.2%

Due to the small amount of data available, the evaluation is performed using a leave-one-out approach. Since AdaBoost has showed the best results in all scenarios, we only indicate the error rates using that classifier. We were able to achieve a APCER of 0.4% at 1.2% BPCER for the histogram features. Using this feature, we observed false positives (FP) especially in cases using the 2D print attacks. In case of the KP feature vector we achieved 4.2% APCER at 7.2% BPCER while having FP mostly at the attacks using the silicone 3D print and the painted latex masks. By combining both feature vectors into a common one and training them with AdaBoost we were able to reduce the APCER to 0% at 0.2%. This is consistent with our assumption that both properties contain complementary information that together allow a meaningful distinction of the classes.

6 Conclusion

We presented an experimental study on evaluating the vulnerability of face recognition system towards presentation attacks. We proposed a novel multispectral face image database comprised of 91 subjects and several face presentation attacks. We explored the intrinsic characteristics of UV and VIS images and used global and local features to quantify the captured images as bona fide or attack. Our results indicate that UV images include useful information for PAD.

Acknowledgment

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

[3] Proportion of attack presentations using the same PAI species incorrectly classified as bona fide presentations in a specific scenario

[4] Proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario

References

[ASJT17] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou; Tzimiropoulos, Georgios: , 3D Face Reconstruction from a Single Image, 2017. http://cvl-demos.cs.nott.ac.uk/vrn/.

[BAO17] Buza, Emir; Akagic, Amila; Omanovic, Samir: Skin detection based on image color segmentation with histogram and K-means clustering. In: 2017 10th International Conference on Electrical and Electronics Engineering (ELECO). IEEE, S. 1181–1186, 2017.

[DD16] Damer, Naser; Dimitrov, Kristiyan: Practical View on Face Presentation Attack Detection. In: BMVC. 2016.

[De12] De Marsico, Maria; Nappi, Michele; Riccio, Daniel; Dugelay, Jean-Luc: Moving face spoofing detection via 3D projective invariants. In: 2012 5th IAPR International Conference on Biometrics (ICB). IEEE, S. 73–78, 2012.

[Fi88] Fitzpatrick TB: The validity and practicality of sun-reactive skin types i through vi. Archives of Dermatology, 124(6):869–871, 1988.

[In16] International Standards Organization: , ISO/IEC 30107-1:2016 - Information technology – Biometric presentation attack detection – Part 1: Framework, 2016.

[Ra17] Raghavendra, R.; Raja, K. B.; Venkatesh, S.; Cheikh, F. A.; Busch, C.: On the vulnerability of extended Multispectral face recognition systems towards presentation attacks. In: 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). S. 1–8, Feb 2017.

[RB17] Ramachandra, Raghavendra; Busch, Christoph: Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Computing Surveys (CSUR), 50(1):8, 2017.

[Ru11] Rublee, Ethan; Rabaud, Vincent; Konolige, Kurt; Bradski, Gary R: ORB: An efficient alternative to SIFT or SURF. In: ICCV. Jgg. 11. Citeseer, S. 2, 2011.

[Sa18] Samatzidis, T.; Siegmund, D.; Goedde, M.; Damer, N.; Braun, A.; Kuijper, A.: The Dark Side of the Face: Exploring the Ultraviolet Spectrum for Face Biometrics. In: 2018 International Conference on Biometrics (ICB). S. 182–189, Feb 2018.

[SKJ16] Steiner, H.; Kolb, A.; Jung, N.: Reliable face anti-spoofing using multispectral SWIR imaging. In: 2016 International Conference on Biometrics (ICB). S. 1–8, June 2016.

[Zh16] Zhang, Kaipeng; Zhang, Zhanpeng; Li, Zhifeng; Qiao, Yu: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 215

Privacy Evaluation Protocols for the Evaluation of Soft-Biometric Privacy-Enhancing Technologies

Philipp Terhörst[1,2], Marco Huber[1], Naser Damer[1,2], Peter Rot[3], Florian Kirchbuchner[1], Vitomir Struc[3], Arjan Kuijper[1,2]

Abstract: Biometric data includes privacy-sensitive information, such as soft-biometrics. Soft-biometric privacy enhancing technologies aim at limiting the possibility of deducing such information. Previous works proposed several solutions to this problem using several different evaluation processes, metrics, and attack scenarios. The absence of a standardized evaluation protocol makes a meaningful comparison of these solutions difficult. In this work, we propose privacy evaluation protocols (PEPs) for privacy-enhancing technologies (PETs) dealing with soft-biometric privacy. Our framework evaluates PETs in the most critical scenario of an attacker that knows and adapts to the systems privacy-mechanism. Moreover, our PEPs differentiate between PET of learning-based or training-free nature. To ensure that our protocol meets the highest standards in both cases, it is based on Kerckhoffs's principle of cryptography.

Keywords: Face, soft-biometric privacy, privacy-enhancing technologies, evaluation protocols.

1 Introduction

Recent works on soft-biometrics showed that privacy-sensitive information, such as gender, age, ethnicity, or even health can be deducted from biometric data of an individual [DER16, Te19c]. However, for many applications, biometric data is expected to be used for recognition purposes only, and extracting such information without the user's agreement raises major privacy issues [Ki13]. Consequently, this kind of data is given special protection, e.g. by the European Union with its General Data Protection Regulation [CotEU16]. Soft-biometric privacy aims at suppressing this privacy-sensitive information in biometric data, to prevent a potential misuse (*function creep)* of this information. Previous works proposed several solutions to this problem. However, since these works consider different evaluation metrics and attack scenarios, a meaningful comparison is difficult. In this work, we propose a standardized framework for evaluating the performance of PETs on soft-biometric privacy. We introduce propose privacy evaluation protocols (PEPs) for learning-based and training-free scenarios. Following the Kerkhoff principle, our PEPs build on the critical scenario of a function creep attacker that knows and adapts to the system's privacy-mechanism. Our PEPs include a detailed description of the data handling, the choice and the training of the attack estimators, as well as, robust and meaningful evaluation metrics for both aspects of soft-biometric privacy, suppressing privacy-risk information and maintaining recognition ability.

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
[3] Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

2 Related Work

Previous works on soft-biometric privacy either proposed solutions at the image-level [Su11, OR14, MR17, Mi18, MRR18, MRR19] or at template-level [MFV19, Te19a, Te19b, Te20b, Te20a]. At the image-level, Suo et al. [Su11] proposed a gender conversion approach that exchanges facial components of a given face with similar components of the opposite gender. Othman and Ross [OR14] proposed an image-based approach by applying face morphing. To disturb the original gender of an individual they morph the individuals‘ image with an image of the opposite gender. In [Ch18], imperceptible noise was used to suppress k attributes at the same time. However, this noise is trained to suppress attributes from only one specific neural network classifier and consequently, may not generalize to other classifiers. Mirjalili and Ross [MR17] iteratively perturb face images until the soft-biometric attribute assigned by arbitrary estimator flips. More recently, Mirjalili et al. [Mi18] used semi-adversarial networks (SAN) to suppress the gender information in images. SANs are auto-encoders with adversarial training that aim to maximize the performance of a face matcher and to minimize the performance of an estimator. In [MRR18] and [MRR19], the authors extended the idea of SANs to (a) an ensemble of SANs and (b) combining a diverse set of SAN models to compensate for each other's weaknesses.

Recently, template-based solutions received a lot of attention. In 2019, Terhörst et al. [Te19b] proposed similarity-sensitive noise transformations to suppress privacy-sensitive attributes in face representations in an unsupervised manner. Concurrently, Morales et al. [MFV19] introduced SensitiveNets, a network that suppresses target information in face templates based on triplet loss learning. In [Te19a], Terhörst et al. proposed Incremental Variable Elimination (IVE). IVE iteratively determines the most privacy-risk variables and deletes them from the face template. Bortolato et al. [Bo20] proposed PFRNet, a face template learning framework that disentangles identity from soft-biometrics to enhance privacy. In [Te20b], Terhörst et al. proposed Negative Face Recognition (NFR). This unsupervised approach stores only complementary identity information to enhance the user's privacy. Exploting the structural differences between face recognition (use-case) and facial attribute estimation (attack scenario), same authors proposed a privacy-preserving face recognition approach based on minimal information units (PE-MIU) [Te20a].

The following list summarizes the limitations of previous works and demonstrates the need for a standardized evaluation protocol:

- **Violation of Kerkhoffs‘ principle**: Most previous works [Su11, OR14, MR17, Mi18, MRR18, MRR19, Ch18] assume an attacker with only restricted resources and knowledge about the systems privacy mechanism.
- **Gender focus**: Most previous works focus mostly on the evaluation of the binary characteristic gender. The effectiveness of categorical or continuous attributes, such as race and age, is not well investigated [Su11, OR14, MR17, Mi18, MRR18, MRR19].
- **Non-robust evaluation metrics:**: Evaluation metrics (accuracy) used in most previous works [MFV19, OR14, Ch18] are sensitive to the underlying data distribution and thus, vulnerable to unbalanced data.
- **Non-standardized evaluation process**: Due to no established evaluation protocols, a meaningful comparison of PETs is difficult.

3 Attack Scenario

For the privacy evaluation protocol, we assume the following attack scenario: the attacker gained unauthorized access to the face templates or images (stored or transmitted) used to recognize individuals. The attacker may have extensive knowledge of how these were created and what method was used to enhance the privacy of the users. Moreover, the attacker may have access to computational power and an annotated face dataset. Accordingly, we follow Kerckhoffs's principle known from cryptography, which Shannon formulated as *„the enemy knows the system being used"* [Sh49]. The attacker's objective is the function creep of the privacy-sensitive information of the individuals for an unknown purpose.

4 Framework / Protocol

In this section, we propose three soft-biometric *privacy evaluation protocols (PEPs)*. We distinguish between the evaluation of training-free (PEP-TF) and learning-based (PEP-LB) PETs. The learning-based PETs need additional data about the suppressed attributes for the training. PEP-TF requires no additional training and can be directly applied to the data. For the learning-based scenario, we suggest an additional (third) loose protocol (PEP-LBL) if the amount of data is not sufficient to perform the strict evaluation protocol (PEP-LBS).

4.1 Preliminary

The first step of the protocol is to split the data set in approximately equally sized folds k with $k \geq 3$. This split should preserve the statistical distribution of the data set and enforce subject-exclusiveness. This means that images of an individual are not distributed over multiple folds but only included in one fold exclusively. This is done to ensure that virtual attackers learn abstract soft-biometric information and do not rely on learned identity information when predicting soft-biometric attributes. The folds are used to perform k-fold cross-validation. The number of folds used for training, development (parameter tuning), and testing are specified in an extended notation: PEP-LBS-N_{train}-N_{dev}-N_{test}. The N values indicate the number of folds for the specific step. For instance, PEP-LBS-2-1-2 would indicate that the learning-based and loose protocol was performed with two folds as training set, one fold for hyperparameter-tuning, and two folds for testing. After splitting the data in the different folds, the feature vectors are scaled to unit-length and further normalized. Feature normalization, such as z-score or min-max scaling, is applied in the same way as the protocol presented below. These two steps ensure a meaningful start for the attack estimators.

4.2 PEP-LBS: Learning-based and Strict Evaluation Protocol

The *learning-based and strict privacy-enhancing protocol (PEP-**LBS**)* assures that the same data is not used multiple times during the evaluation process. The protocol assumes that the PET includes a training process. Therefore, the original data set is divided into three parts D_{train}, D_{dev}, and D_{test} (which all may consist of multiple folds). The D_{train} set is used to train the PET and the D_{dev} to fine-tune possible hyper-parameters of the method. The D_{test} is transformed using the trained and fine-tuned privacy-enhancing method and further divided into the three subsets: T_{train}, T_{dev}, and T_{test}. It is important to note that T_{train},

T_{dev}, and T_{test} are subsets of the transformed D_{test} and not the transformed D_{train} and D_{dev}. T_{train} is then used to train the different FCEs. T_{dev} is used to fine-tune the hyper-parameters of these FCEs. The T_{test} set is used to evaluate the performance of the PETs in regard of its recognition performance and the suppression performance on the FCEs. A schematic view of the PEP-LBS protocol can be seen in Figure 1a. When using the PEP-LBS we recommend to choose the number of folds in the test subset, $N_{test} \geq 3$.

(a) PEP-LBS (b) PEP-LBL

Fig. 1: Schematic of the data handling of both learning-based protocols PEP-LB.

4.3 PEP-LBL: Learning-based and Loose Evaluation Protocol

In PEP-LBS, dividing the D_{test} into T_{train}, T_{dev}, and T_{test}, requires an appropriate large test set D_{test} and thus, a large amount of data. Since this is often not available, we introduce the *learning-based and loose protocol (PEP-**LBL**)*. In this protocol, the data separation is loosened. This comes at the cost of a partial overfit of the PET on T. The D_{train} subset is used to train the privacy-enhancing method. The D_{dev} subset is used to adjust the hyper-parameters of the PET. Afterwards, all three subsets D_{train}, D_{dev}, and D_{test} are transformed using the PET into T_{train}, T_{dev}, and T_{test}. T_{train} is used to train the estimators of the attacker and T_{dev} to fine-tune the parameters of the estimators. T_{test} is only used to evaluate the PET. The loose protocol provides a trade-off if splitting the test set D_{test} to evaluate the FCEs would lead to too small subsets that meaningful results cannot be obtained. To prevent this, the train set D_{train} and the development set D_{dev} are used twice, once in their unaltered templates/images to train and fine-tune the PET and once in their transformed ones T_{train} and T_{dev} to train and fine-tune the attack estimators. A schematic view of the PEP-LBL protocol is shown in Figure 1b.

4.4 PEP-TF: Training-free Evaluation Protocol

The proposed *training-free evaluation protocol (PEP-**TF**)* assumes that the PET does not require a training phase. Therefore, the three parts of the original data set, D_{train}, D_{dev}, and D_{test} are directly transformed by the PET to the modified templates/images T_{train}, T_{dev}, and T_{test}. T_{train} is used to train the different FCEs, T_{dev} is used to adjust the hyper-parameter of those estimator models and T_{test} is then used to evaluate the performance of the privacy-enhancing method. An illustration of the PEP-TF protocol is shown in Figure 2.

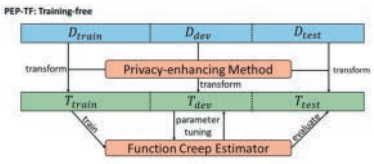


Fig. 2: Illustration of the data handling for the training-free protocol PEP-TF.

4.5 Function Creep Estimators

In the proposed attack scenario, the function creep attacker deploy *function creep estimators (FCEs)* to determine privacy-sensitive attributes that were previously obscured by the transformation through the PET used. These FCEs are trained and fine-tuned as described in the used protocol. The hyperparameter tuning can be done, for example, via Grid Search, Random Search or Bayesian Optimization. In Section 4.1, the pre-processing of the data was already described.

The template-based approaches are evaluated using the extracted feature representations of the face images. For the possible FCEs, we recommend well-known estimators that should be used as a baseline to assess the quality of the PETs. These include *random forest*, *support-vector machines*, *k-nearest neighbors* and *logistic regression*. This choice is based on (a) their membership to different kinds of machine learning models and (b) the fact that these perform evidently well on face templates [Te20b, Te19a]. Each FCE is independently trained twice: first, on the unmodified data and second, on the transformed data that was modified by the PET. This allows us to compare the performance of the estimators without having noise due to different test samples. The training of several different estimators is intended to ensure the robustness of the PET for different kind of attacks. Please note that another attack scenario might come from regenerating a face image from a template and manually investigating this. However, patterns of privacy-sensitive information in templates are generally easily detectable due to the feature entanglement during the learning process.

In contrast to PET based on template-level, image-based approaches have to deal with large-scale and more restricted feature spaces. Image-based approaches have the advantage that, for many attributes, the modified representations can be evaluated by humans as well. However, the choice of function creep estimators should additionally include machine-based solutions since these solutions might catch suspicious artifacts that humans are not aware of. Due to the large-scale nature of images, (a) CNN approaches [KSH12] should be used as potential FCEs or (b) a combination of lower-dimensional handcrafted features, such as LBPH [AHP06], with the proposed template-based estimators.

5 Evaluation

So far, the protocol descriptions focus on the data handling and the training of PETs and FCEs. Based on this, this section describes how the PETs can be robustly evaluated in regard to the FCEs. The challenge of soft-biometric privacy describes a trade-off between maintaining the recognition performance of face representations and suppressing the predictability of privacy-sensitive attributes within these. To evaluate both aspects of the trade-off, the attribute estimation and the recognition results of the modified and unmodified face representations are compared. For the evaluation of the attribute suppression performance, the predictions of the FCEs on the un/modified representations of T_{test} are used. The evaluation of the recognition performance is based on the un/modified representations of T_{text}.

The recognition performance is the most important factor of recognition systems, since it measures its major purpose. We recommend to evaluate these in terms of *receiver operating characteristic* (ROC) curves with *false non-match rates* (FNMR) at a different *false*

match rates (FMR) as defined in the ISO standard [ISO]. ROC curves provide a broad performance overview independent of the application and allow to compare the recognition performance of the unmodified baseline with the PETs. For more specific comparisons, FNMR at a fixed FMR of 10^{-3} or smaller can be used as recommended by the European Border Guard Agency Frontex [Fr17].

To evaluate the suppression performance of PETs, we recommend the use of the *balanced accuracy*. This balanced accuracy is equivalent to the standard accuracy definition with class-balanced sample weights. This allows an unbiased performance measure on testing data with unbalanced attribute information. The suppression performance of PETs can be compared by providing the attribute estimation performance of the unmodified baseline and compare these with the estimation performances of the FCEs on the PET-modified representations. For a single value comparison on the suppression performance the *suppression rate* [Te20b] can be used. The suppression rate $\gamma = \frac{acc_{org} - acc_{mod}}{acc_{org}}$ is defined by the difference in prediction accuracy between unmodified (acc_{org}) and PET-modified (acc_{mod}) representations.

To measure the benefit of applying a PET, the *privacy gain identity loss coefficient (PIC)* [Te19b] is a suitable tool. The PIC is defined as $PIC = \frac{AE' - AE}{AE} - \frac{RE' - RE}{RE}$ where AE and AE' denote the attribute prediction errors of an FCE. RE and RE' define the recognition errors with and without the privacy-enhancement of the face representations. In Figure 3, equipotential lines for different PIC-values are shown and visualize the trade-off. The PIC values the relative error of the FCE prediction with the relative error of the recognition performance. Consequently, it directly measures the benefit of using the PET such that a higher coefficient states a higher benefit.

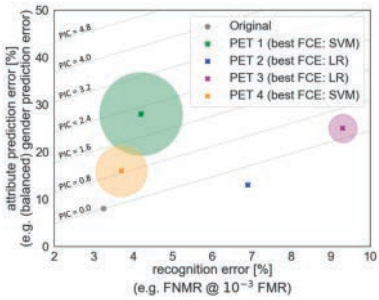


Fig. 3: Example of a recognition-attribute plot [Bo20]. The attribute prediction error is shown over the recognition error for the unmodified baseline and the different PETs. The attribute error refers to the most successful FCE. The size of the shaded areas refer to the PIC coefficient for a PET. Additionally, equipotential lines for different PIC-values are shown in grey.

To visualize the worst-case privacy-enhancing performance, we recommend the use of recognition-attribute plots [Bo20], as shown in Figure 3. This plot shows the recognition error (e.g. the FNMR at 10^{-3}FMR) over the balanced prediction error of an attribute (e.g. gender). The attribute prediction error refers to the most successful FCE, to simulate the most critical attack scenario. In the plot, the unmodified baseline is shown, as well as the PETs under the specification of the most successful FCE. This allows a complete evaluation of the trade-off between suppressing an attribute and maintaining the recognition performance. To further visualize the benefit of applying a PET, the size of the shaded areas around a PET represents its PIC coefficient.

6 Conclusion

Extracting privacy-sensitive information, such as demographics or health information, about an individual from biometric data without consent is considered a major privacy issue. Recent works proposed PETs under different evaluation processes, metrics, and considered attack scenarios. This makes a meaningful comparison of these methods challenging. To enhance the comparability of PETs, and thus enhance the development of this field, we propose PEPs in the most critical attack scenario of a function creep attacker that knows and adapts to the systems privacy-mechanism. We propose three PEPs to ensure sufficient use of the data concerning the nature of the evaluated PET. This includes efficient and independent data handling, training of PETs and FCEs, and robust evaluation metrics for both aspects of soft-biometric privacy.

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

References

[AHP06] Ahonen, Timo; Hadid, Abdenour; Pietikainen, Matti: Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Trans. Pattern Anal. Mach. Intell., 28(12):2037–2041, December 2006.

[Bo20] Bortolato, B.; Ivanovska, M.; Rot, P.; Krizaj, J.; Terhörst, P.; Damer, N.; Peer, P.; Struc, V.: Learning Privacy-Enhancing Face Representations through Feature Disentanglement. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG). IEEE Computer Society, Los Alamitos, CA, USA, pp. 45–52, may 2020.

[Ch18] Chhabra, Saheb; Singh, Richa; Vatsa, Mayank; Gupta, Gaurav: Anonymizing k Facial Attributes via Adversarial Perturbations. In (Lang, Jérôme, ed.): Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org, pp. 656–662, 2018.

[CotEU16] Council of the European Union, European Parliament: , Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[DER16] Dantcheva, Antitza; Elia, Petros; Ross, Arun: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE Trans. Information Forensics and Security, 11(3):441–467, 2016.

[Fr17] Frontex: Best Practice Technical Guidelines for Automated Border Control (ABC) Systems. 2017.

[ISO] : Information technology - Biometric performance testing and reporting - Part 1: Principles and framework. Standard, International Organization for Standardization.

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

[KSH12] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E: ImageNet Classification with Deep Convolutional Neural Networks. In (Pereira, F.; Burges, C. J. C.; Bottou, L.; Weinberger, K. Q., eds): Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc., 2012.

[MFV19] Morales, Aythami; Fiérrez, Julian; Vera-Rodríguez, Rubén: SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. CoRR, abs/1902.00334, 2019.

[Mi18] Mirjalili, Vahid; Raschka, Sebastian; Namboodiri, Anoop M.; Ross, Arun: Semi-adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images. In: 2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018. IEEE, pp. 82–89, 2018.

[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 2017, Denver, CO, USA, October 1-4, 2017. IEEE, pp. 564–573, 2017.

[MRR18] Mirjalili, Vahid; Raschka, Sebastian; Ross, Arun: Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers. In: 9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018. IEEE, pp. 1–10, 2018.

[MRR19] Mirjalili, Vahid; Raschka, Sebastian; Ross, Arun: FlowSAN: Privacy-Enhancing Semi-Adversarial Networks to Confound Arbitrary Face-Based Gender Classifiers. IEEE Access, 7:99735–99745, 2019.

[OR14] Othman, Asem A.; 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 - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II. volume 8926 of Lecture Notes in Computer Science. Springer, pp. 682–696, 2014.

[Sh49] Shannon, Claude E: Communication theory of secrecy systems. Bell system technical journal, 28(4):656–715, 1949.

[Su11] Suo, Jin-Li; Lin, Liang; Shan, Shiguang; Chen, Xilin; Gao, Wen: High-Resolution Face Fusion for Gender Conversion. IEEE Trans. Systems, Man, and Cybernetics, Part A, 41(2):226–237, 2011.

[Te19a] Terhörst, Philipp; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Suppressing Gender and Age in Face Templates Using Incremental Variable Elimination. In: 2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019. IEEE, pp. 1–8, 2019.

[Te19b] Terhörst, Philipp; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. Appl. Intell., 49(8):3043–3060, 2019.

[Te19c] Terhörst, Philipp; Huber, Marco; Kolf, Jan Niklas; Zelch, Ines; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Reliable Age and Gender Estimation from Face Images: Stating the Confidence of Model Predictions. In: 10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, Florida, USA, September 23-26, 2019. IEEE, 2019.

[Te20a] Terhörst, Philipp; Riehl, Kevin; Damer, Naser; Rot, Peter; Bortolato, Blaz; Kirchbuchner, Florian; Struc, Vitomir; Kuijper, Arjan: PE-MIU: A Training-Free Privacy-Enhancing Face Recognition Approach Based on Minimum Information Units. IEEE Access, 8:93635–93647, 2020.

[Te20b] Terhörst, Philipp; Huber, Marco; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Unsupervised Enhancement of Soft-biometric Privacy with Negative Face Recognition. CoRR, abs/2002.09181, 2020.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 223

Minutiae-based Finger Vein Recognition Evaluated with Fingerprint Comparison Software

Katy Castillo-Rosado[1,0], Michael Linortner[2,0], Andreas Uhl[2], Heydi Mendez-Vasquez[1], José Hernandez-Palancar[1]

Abstract: Finger vein recognition is a biometric authentication technique based on the vein patterns of human fingers. Despite the fact that classical approaches are based on correlation, the topology of vein patterns allows the use of minutiae points for their representation. Minutiae points are the most used features for representing ridge patterns in fingerprints. In literature, it has been shown that minutiae can be used for finger vein comparison, but low image quality provokes that many spurious minutiae are extracted from them. In this work, a preprocessing method is presented, that combines classical digital image processing methods and level set theory in order to extract a set with the most reliable minutiae. The experiments were performed on two publicly available databases and different comparison methods were used for testing the representative character of the minutiae set extracted. The results showed that even though the amount of extracted minutiae is around 15-30, effective identification is possible.

Keywords: Finger veins, minutiae, recognition.

1 Introduction

Vascular pattern recognition, also called vein recognition, utilizing blood vessels located underneath the skin of a finger, hand, or wrist as a biometric trait has become an emerging technology in the field of biometrics. Under near-infrared (NIR) light the veins appear as dark structure which is captured on gray scale images with a NIR sensitive camera. The depicted vein structure is assumed to be unique for an individual, even for each hand or finger of a person [Uh20].

The veins are segmented using different methods, like *principal curvature* (PC) [Ch09], resulting in a binary vein pattern which is used as biometric template. The classical approach to compare two such templates is to obtain a similarity score by applying correlation. Another approach is to utilize intersection, branching or endpoints of the extracted vein pattern analogous to minutia points used in fingerprint recognition. Using minutiae points in vein recognition has already been investigated in literature. In [Yu09] and [WLC08] minutiae points are extracted from finger and hand vein images, respectively. For comparison of the biometric templates *modified Hausdorff distance* (MHD) is utilized in both cases. A similarity score between two minutiae sets from hands is calculated in [Ur11] by counting corresponding minutiae pairs that have similar relative positions and angles. In [Li14]

[1] Advanced Technologies Application Center (CENATAV), Biometrics, 7ma A ♯21406 *b*/ 214 and 216, Siboney, Playa, P.C. 12200. Havana, Cuba, {krosado,hmendez,jpalancar}@cenatav.co.cu
[2] Department of Computer Sciences, University of Salzburg, Salzburg, Austria, {mlinortner,uhl}@cs.sbg.ac.at
[0] These authors contributed equally

minutiae pairing is done using SVD and the comparison is based on an average similarity degree after false pairs have been removed by applying a LBP variant locally. *Minutiae cylinder codes* (MCC) applied on hand vein minutiae points is described in [HTB13]. MCC has been originally introduced in fingerprint recognition [CFM10a].
The main contribution of this work is to show if it is possible to apply classical fingerprint minutiae comparison software on minutiae points extracted from vein images to achieve a high performance in accuracy and speed. To the authors best knowledge this has not been investigated in literature so far. For vein segmentation a novel method inspired by the general fingerprint minutiae extraction process is introduced, where some well known techniques for image processing with methods from level set theory are combined. From the experiments it can be seen, that despite the fact that the extracted minutiae set is small (25-30) it can correctly identify the vein pattern.
An advantage applying fingerprint minutiae comparison technology on vein images is, that there can be already existing solutions adopted to finger vein recognition, like for embedded systems or so called *Match-on-Card* (MoC) systems [BSV06], which would then be the first MoC system in finger vein recognition.
Four classical minutiae-based comparison software are evaluated, see section 3.

2 Proposed feature extraction

Inspired by the general idea for preprocessing fingerprint images with low quality, the following method is presented for extracting minutiae from finger vein images. The main idea is to extract the most reliable minutiae points. The method combines a group of well known techniques from digital image processing with methods from level set theory. For this, the vein image can be modeled as a fluid interface. Applications of fluid interfaces include breaking surface waves, in which factors such as topological connectivity and boundary conditions play significant roles [SS03].
The input image I is treated as a surface where each pixel is replaced by the curvature of the surface at that point. To calculate the curvature, a method based on level set theory [OS88] is used. This method has a positive impact on computational methods for surface movements and has been used to solve a wide kind of problems. In this method, the mean curvature δ of an image I is computed as $\delta = \nabla \cdot n$, where ∇ is the gradient operator and n is the outward drawn normal [Sm03] defined as $n = \frac{\nabla_I}{|\nabla_I|}$.
By resting in this theory for detecting the ridges and valleys location, and by using well known methods from literature [GW06] for enhancing the image, the method for extracting minutiae points is introduced and it is explained below. A visual example of the method output is presented in figure 1.
Due to the low quality presented in vein image, first the image has to be smoothed. In order to reduce the pixelation effect during capturing, the mask size which must to be used in this case is small. The images usually present very low contrast in all the dimensions. Even in some cases, there is not a visual difference between finger veins and the background. Therefore, a process of image normalization is needed. After normalizing the image, it will be distorted. This process may cause the appearance of some spurious artifacts. An oriented contextual filtering can enhance the image and it can highlight the truly vein pattern structure. The orientation field is calculated by using the classic gradient method from

literature [Ma09], in this case, the window (w_θ) for processing the image needs to be selected from a small set of possible values. The filtering stage is done with a bank of 120 low-pass oriented filters.
The binarization process is not trivial. In fingerprints, ridges and valleys have similar dimensions, and the threshold can be determined by local average. For vein images, this process is not appropriate, because ridges and valleys have different width. So, it is necessary to calculate the highest values for ridges and the lowest values for valleys. For this purpose, by relying on level set theory, the normalized gradient divergence is calculated, in order to highlight the maximum and minimum peaks in the image. Maximum peaks have a positive value, minimum peaks have a negative value and the ones that are not peaks have values close to 0. For identifying these peaks, Otsu thresholding is applied to the absolute value of the normalized gradient divergence. Then, these peaks are used to estimate a reliable local threshold for binarizing the image. For this purpose, the size of the windows for smoothing the image before calculating the divergence (w_δ) and for binarizing the image (w_β) need to be estimated. After binarization, the skeleton image is calculated and minutiae are extracted with well known methods from literature [ZS84]. Only bifurcations are selected, because endings are very probably spurious minutiae. For calculating a reliable minutiae direction, bifurcations where the length of at least one of their branches is less than a certain threshold (γ_{min}) are eliminated. Also, the length of each branch line (λ) where the trace is going to stop for calculating direction needs to be declared. Direction field for each minutiae is calculated in the same way as for fingerprint minutiae.

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

Fig. 1: Example image from the UTFVP database showing the main steps of the method process. (a) The enhanced image with the oriented contextual filter. (b) The normalized gradient divergence. (c) The binarized image. (d) Vein skeleton with the extracted minutiae.

3 Experiments

The experiment's main purpose is to show that finger vein patterns can be correctly described by a small minutiae set, corroborating in this way, the idea of using minutiae points for finger vein recognition.
The following four fingerprint minutiae comparison software packages are used: The publicly available Bozorth3 as a part of the NIST Biometric Image Software (NBIS) Release 5.0.0 and the Minutiae Cylinder Code (MCC) SDK [Ca10, CFM10b], as well the IDKit SKD Version 9.0 from Innovatrics and the VeriFinger 11.2 Extended SDK from Neurotechnology, two commercial products. Latter both offer a MoC system which is NIST Minutiae Interoperability Echange (MINEX) compliant. MHD is used as an additional point based comparison method.
To compare the minutiae-based approaches to classical vein recognition techniques, PC has been chosen, which shows a good baseline performance in vein recognition. PC is a vein segmentation method and cross-correlation is used to obtain a similarity score between two templates [Ch09].

IDKit and VeriFinger require the minutiae input in the ANSI/INCITS 378-2004 [AN04] binary format, MCC can read the minutiae from ASCII files. All three use the information of image size and resolution. Some tests showed, that different comparison scores are produced when the resolution value changes, which is true for all three software packages. Therefore, a resolution value is chosen such that the vein image ROI has approximately the same size as the bounding box of a fingerprint with a resolution of 500 dpi. For the UTFVP ROIs (336×128 pixels) a resolution of 134 pixel per centimeter (pcc) and for the PLUS ROIs (368×96 pixels) a resolution of 147 pcc is derived.
The proposed preprocessing method was implemented in Matlab 2019. MHD and PC as well the comparison score evaluation have been implemented in C++ using the OpenCV library version 3.4.2.

3.1 Data sets

Two publicly available finger vein data bases are used for the experiments:

- The *University of Twente Finger Vascular Pattern Database* (UTFVP)[TV13] contains images of six fingers (index, middle and ring finger) of both hands of 60 subjects. Four samples of each finger have been acquired.
- The *PLUSVein-FV3 Finger Vein Database* [KPU18] (PLUS) contains 4 data sets of the same subjects and fingers with images captured from the dorsal (D) and palmar (P) view, both acquired with LED and Laser illumination, respectively. It also provides already extracted ROI images which were used in the experiments.

3.2 Parameter search

The quality of the extracted features highly depends on the image quality. Therefore, the parameter selection for the method is crucial. Databases have different acquisition characteristics and quality of the obtained data varies from one database to another. Therefore, a parameter search is needed for obtaining the best possible results.
To avoid over fitting a 2-fold validation is employed, splitting the data in such a way, that each fold contains the images from half of the subjects which are assigned to each fold randomly. For each parameter set the *equal error rate* (EER) is calculated for each fold. Based on the lowest EER value found in fold 1, the comparison scores of a parameter set in fold 2 are selected and vice versa, then they are combined and on them again the EER is evaluated and reported. This is done 100 times, each time splitting the subjects randomly.
For the preprocessing and feature extraction method the parameter value search is performed on a total of 5 parameters. The possible values for each of these parameters are selected by studying each step involved in the entire process. The combinations of all parameters make a total of 144 different settings: $w_\theta \in \{8,16,32\}$, $w_\delta \in \{4,8,12,16\}$, $w_\beta \in \{32,40,64\}$, $\gamma_{min} \in \{10,20\}$ and $\lambda \in \{10,20\}$.

In case of the MCC SDK also a parameter search is necessary. After investigating the behavior of different MCC settings for fingerprint tenprint impressions and fingerprint latent

impressions [Ca10, CFM10b], some parameters were set based on previous researches for latent fingerprints [VMM19] and for some of them a set of possible values was selected. Due to the high amount of possible parameter settings, first a parameter search was done only for one minutiae set of the UTFVP database (total of 1527 parameter combinations). Then the parameters of the first 50 best results have been investigated and a subset has been chosen for enrollment: $radius = 300, \sigma_s = 9, \sigma_d = \frac{\pi}{6}, \mu_\psi = 0.002, \omega = 100, Min_{VC} = 0.03, Min_M = 1$ and for comparison: $Min_{ME} = 0.03, \delta_t = \frac{\pi}{6}, \mu_P = 32, \mu_{\rho 1} \in \{\frac{1}{30}, \frac{1}{24}\}, \tau_{\rho 1} \in \{-50, -100\}, \mu_{\rho 2} \in \{\frac{\pi}{8}, \frac{\pi}{4}\}, \tau_{\rho 2} = -25, \mu_{\rho 3} = \frac{\pi}{16}, \tau_{\rho 3} \in \{-28, -40\}, nrel = 4$.

3.3 Evaluation

Evaluation is done applying the fingerprint verification competition 2006 [3] (FVC) protocol and each finger of an individual is considered as a single class. Next to the recognition accuracy, execution times for the template comparison have been evaluated, to show differences in time performance between minutiae-based and the classical correlation-based finger vein recognition methods. All templates are loaded into RAM so that only template comparison time is considered. The time needed to execute 66780 comparisons has been measured (UTFVP dataset following FVC protocol).

3.4 Results

Table 1 - 5 report the recognition performance for the used data sets and different methods applied on them. For all reported values the average (avg) as well the standard deviation (std) have been calculated using the results of the $N = 100$ different evaluations as described in section 3.2. For the EER additionally the minimum and maximum value of the 100 fold splits is presented to show how much the performance can vary depending on the selection of subjects for the parameter estimation. Beside EER, the ZeroFMR, FMR100, ZeroFNMR, FNMR100 and *area under curve* (AUC) are reported.
In general, the four minutiae-based comparison software SDKs Bozorth3, IDKit, VeriFinger and MCC perform quite similar, being VeriFinger the one which obtains the best results. On the PLUS data sets MCC lies very close to VeriFinger and IDKit, while on the UTFVP database its performance is a little bit lower and is similar to Bozorth3. MHD, which is a more naive approach, clearly shows the lowest performance on all data sets.
The classical correlation-based PC shows a better performance on all databases.
The results indicate, that on the dorsal view of a finger a better performance can be achieved. It seems there are more structures visible which are feasible for minutiae extraction. Table 6 shows the evaluation of the execution times for 66780 comparisons, averaged over 10 runs. It clearly shows that all minutiae-based approaches run almost two order of magnitude faster than the classical correlation-based PC approach.

[3] http://bias.csr.unibo.it/fvc2006/perfeval.asp

Tab. 1: Recognition performance for the UTFVP data set.

UTFVP	EER in%				ZeroFMR in%		FMR100 in%		ZeroFNMR in%		FNMR100 in%		AUC in%	
	avg	std	min	max	avg	std	avg	std	avg	std	avg	std	avg	std
Bozorth3	12.1	0.4	11.5	13.4	57.3	5.8	25.2	0.8	91.1	1.4	89.6	1.3	93.85	0.33
IDKit	10.1	0.3	9.6	10.9	47.8	8.6	19.0	0.6	98.8	0.1	77.9	2.7	95.49	0.19
VeriFinger	6.8	0.2	6.4	7.2	39.0	4.2	15.5	0.5	11.0	0.1	11.0	0.1	96.41	0.29
MHD	14.7	0.4	13.8	15.8	72.5	5.0	34.4	1.2	99.7	0.3	87.3	2.3	92.50	0.30
MCC	12.2	0.4	11.6	13.4	51.8	5.3	22.9	0.8	99.3	0.5	88.0	2.5	93.88	0.28
PC	0.4	0.1	0.2	0.8	1.2	0.5	0.3	0.1	85.0	10.2	0.0	0.1	99.86	0.05

Tab. 2: Recognition performance for the PLUS-Las-P data set.

PLUS-Las-P	EER in%				ZeroFMR in%		FMR100 in%		ZeroFNMR in%		FNMR100 in%		AUC in%	
	avg	std	min	max	avg	std	avg	std	avg	std	avg	std	avg	std
Bozorth3	12.4	0.3	11.8	13.2	63.9	10.8	22.2	0.5	17.3	0.8	17.3	0.8	88.72	0.54
IDKit	8.2	0.3	7.7	9.1	58.1	7.5	14.2	0.5	99.4	0.0	83.6	2.3	96.25	0.20
VeriFinger	6.5	0.2	6.0	7.0	58.0	10.1	13.0	0.8	9.1	0.2	9.1	0.2	95.01	0.37
MHD	25.8	32.4	10.9	100.0	71.3	13.9	39.3	26.6	99.9	0.1	90.8	4.5	79.32	34.68
MCC	9.2	0.2	8.7	9.7	50.3	9.0	16.1	0.5	99.9	0.1	87.1	2.7	95.74	0.15
PC	1.3	0.3	1.0	2.5	3.9	1.1	1.3	0.4	92.1	7.4	4.5	6.1	99.67	0.13

Tab. 3: Recognition performance for the PLUS-Las-D data set.

PLUS-Las-D	EER in%				ZeroFMR in%		FMR100 in%		ZeroFNMR in%		FNMR100 in%		AUC in%	
	avg	std	min	max	avg	std	avg	std	avg	std	avg	std	avg	std
Bozorth3	7.5	0.3	6.9	8.6	36.1	7.5	12.2	0.7	25.9	2.9	25.9	2.9	94.93	0.35
IDKit	3.7	0.2	3.2	4.3	36.2	7.7	5.3	0.4	99.7	0.9	44.5	4.3	98.80	0.09
VeriFinger	3.1	0.1	2.9	3.5	23.8	4.7	4.4	0.3	9.0	0.1	9.0	0.1	98.13	0.14
MHD	9.8	3.8	0.1	16.7	56.2	13.1	20.4	7.0	100.0	0.1	77.9	14.8	95.39	2.73
MCC	3.9	0.2	3.4	4.6	23.1	4.4	5.6	0.4	100.0	0.0	54.0	6.6	98.58	0.13
PC	0.4	0.1	0.2	0.8	1.7	0.7	0.3	0.1	43.6	34.5	0.0	0.0	99.97	0.02

4 Conclusion

In this work a novel method for extracting minutiae points from finger vein patterns is introduced. This method uses techniques from level set theory for detecting the reliable ridges and valleys from the vein pattern. For evaluating the extraction process the performance of some classical minutiae-based fingerprint comparison techniques are reported. Two state of the art commercial and two publicly available comparison software packages were used. The results achieved by the minutiae-based comparison techniques show promising performances. Although the classical correlation-based PC approach obtains the best results in terms of recognition accuracy, when it comes to the processing time of the comparisons it clearly shows its weakness compared to the minutiae-based comparison methods. The experiments show that there is a trade-off between accuracy and comparison speed using minutiae-based approaches or classical techniques. In future work the binarization process should be enhanced. In this way, more reliable ridges can be detected. Therefore, less minutiae will be missed and less spurious minutiae will be extracted.

Tab. 4: Recognition performance for the PLUS-Led-P data set.

PLUS-Led-P	EER in%				ZeroFMR in%		FMR100 in%		ZeroFNMR in%		FNMR100 in%		AUC in%	
	avg	std	min	max	avg	std	avg	std	avg	std	avg	std	avg	std
Bozorth3	10.6	0.4	10.0	12.4	49.2	5.3	19.8	0.5	23.8	0.8	23.8	0.8	91.60	0.41
IDKit	6.8	0.3	6.4	7.6	40.5	6.4	11.7	0.5	100.0	0.0	71.5	4.2	97.17	0.20
VeriFinger	5.4	0.2	5.0	5.9	35.4	6.9	10.2	0.4	10.1	0.1	10.1	0.1	96.55	0.21
6 MHD	18.4	2.9	10.7	23.7	73.6	7.2	39.7	5.2	100.0	0.1	96.7	3.2	88.43	2.58
MCC	8.4	0.2	7.9	8.9	46.1	6.7	14.8	0.5	100.0	0.0	87.2	2.3	96.16	0.16
PC	0.8	0.1	0.5	1.6	2.8	1.0	0.7	0.2	72.7	24.0	0.4	0.5	99.89	0.05

Tab. 5: Recognition performance for the PLUS-Led-D data set.

PLUS-Led-D	EER in%				ZeroFMR in%		FMR100 in%		ZeroFNMR in%		FNMR100 in%		AUC in%	
	avg	std	min	max	avg	std	avg	std	avg	std	avg	std	avg	std
Bozorth3	7.0	0.4	6.4	8.2	33.0	3.9	11.3	0.6	29.2	1.3	29.2	1.3	95.33	0.41
IDKit	3.7	0.3	3.2	4.9	34.2	12.6	5.6	0.5	96.6	2.5	39.0	4.6	98.91	0.12
VeriFinger	3.2	0.2	2.8	3.8	22.5	5.5	4.9	0.3	9.5	0.2	9.5	0.2	98.22	0.24
MHD	14.4	4.7	7.1	21.6	58.9	8.3	28.1	8.7	99.7	0.6	89.1	11.8	91.66	3.80
MCC	4.3	0.2	3.9	4.9	27.0	6.3	6.4	0.3	99.7	0.4	55.8	4.2	98.46	0.13
PC	0.3	0.1	0.2	0.4	1.0	0.4	0.2	0.1	38.9	29.0	0.0	0.0	99.98	0.01

Tab. 6: Execution times of 66780 template comparisons.

	VeriFinger	MCC	MHD	PC
Time in sec	30.6 ± 0.6	45.7 ± 0.1	17.3 ± 0.2	1678.8 ± 61.7

Acknowledgements

This project was partly funded from the FFG KIRAS project AUTFingerATM under grant No. 864785 and the FWF project ”Advanced Methods and Applications for Fingervein Recognition” under grant No. P 32201-NBL.

References

[AN04] ANSI-INCITS 378-2004: , Information Technology - Finger Minutiae Format for Data Interchange, 2004.

[BSV06] Bistarelli, Stefano; Santini, Francesco; Vaccarelli, Anna: An Asymmetric Fingerprint Matching Algorithm for Java Card TM. Pattern Anal. Appl., 9(4):359–376, Oct. 2006.

[Ca10] Cappelli, Raffaele; Ferrara, Matteo; Maltoni, Davide; Tistarelli, Massimo: MCC: A baseline algorithm for fingerprint verification in FVC-onGoing. In: 11th Int. Conf. on Control, Automation, Robotics and Vision, ICARCV 2010, Singapore, 7-10 December 2010, Proc. pp. 19–23, 2010.

[CFM10a] Cappelli, R.; Ferrara, M.; Maltoni, D.: Minutia Cylinder-Code: A New Representation and Matching Technique for Fingerprint Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 32(12):2128–2141, Dec 2010.

[CFM10b] Cappelli, Raffaele; Ferrara, Matteo; Maltoni, Davide: Minutia Cylinder-Code: A New Representation and Matching Technique for Fingerprint Recognition. IEEE Trans. Pattern Anal. Mach. Intell., 32(12):2128–2141, 2010.

[Ch09] Choi, Joon Hwan; Song, Wonseok; Kim, Taejeong; Lee, Seung-Rae; Kim, Hee Chan: Finger vein extraction using gradient normalization and principal curvature. In (Niel, Kurt S.; Fofi, David, eds): Image Processing: Machine Vision Applications II. volume 7251. Int. Society for Optics and Photonics, SPIE, pp. 359 – 367, 2009.

[GW06] Gonzalez, Rafael C.; Woods, Richard E.: Digital Image Processing (3rd Edition). Prentice-Hall, Inc., USA, 2006.

[HTB13] Hartung, D.; Tistarelli, M.; Busch, C.: Vein minutia cylinder-codes (V-MCC). In: 2013 Int. Conf. on Biometrics (ICB). pp. 1–7, June 2013.

[KPU18] Kauba, Christof; Prommegger, Bernhard; Uhl, Andreas: The Two Sides of the Finger - An Evaluation on the Recognition Performance of Dorsal vs. Palmar Finger-Veins. In: Proc. of the Int. Conf. of the Biometrics Special Interest Group (BIOSIG'18). Darmstadt, Germany, pp. 1–8, 2018.

[Li14] Liu, Fei; Yang, Gongping; Yin, Yilong; Wang, Shuaiqiang: Singular value decomposition based minutiae matching method for finger vein recognition. Neurocomputing, 145:75 – 89, 2014.

[Ma09] Maltoni, Davide; Maio, Dario; Jain, Anil K.; Prabhakar, Salil: Handbook of Fingerprint Recognition. Springer Publishing Company, Incorporated, 2nd edition, 2009.

[OS88] Osher, Stanley; Sethian, James A.: Fronts Propagating with Curvature Dependent Speed: Algorithms Based on Hamilton-Jacobi Formulations. Journal of computational physics, 79(1):12–49, 1988.

[Sm03] Smereka, Peter: Semi-Implicit Level Set Methods for Curvature and Surface Diffusion Motion. Journal of Scientific Computing, 19:439–456, 2003.

[SS03] Sethian, J. A.; Smereka, Peter: Level Set Methods for Fluid Interfaces. Annual Review of Fluid Mechanics, 35(1):341–372, 2003.

[TV13] Ton, B. T.; Veldhuis, R. N. J.: A high quality finger vascular pattern dataset collected using a custom designed capturing device. In: 2013 Int. Conf. on Biometrics (ICB). pp. 1–5, June 2013.

[Uh20] Uhl, Andreas; Busch, Christoph; Marcel, Sebastien; Veldhuis, Raymond: Handbook of Vascular Biometrics. Advances in Computer Vision and Pattern Rec. Springer Nature Switzerland AG, Cham, Switzerland, 2020.

[Ur11] Uriarte-Antonio, J.; Hartung, D.; Pascual, J. E. S.; Sanchez-Reillo, R.: Vascular biometrics based on a minutiae extraction approach. In: 2011 Carnahan Conf. on Security Technology. pp. 1–7, Oct 2011.

[VMM19] Valdes-Ramirez, Danilo; Medina-Pérez, Miguel Angel; Monroy, Raúl: Stacking Fingerprint Matching Algorithms for Latent Fingerprint Identification. In: Progress in Pattern Rec., Image Analysis, Computer Vision, and Applications - 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, Oct. 28-31, 2019, Proc. pp. 230–240, 2019.

[WLC08] Wang, Lingyu; Leedham, Graham; Cho, David Siu-Yeung: Minutiae feature analysis for infrared hand vein pattern biometrics. Pattern Rec., 41(3):920 – 929, 2008. Part Special issue: Feature Generation and Machine Learning for Robust Multimodal Biometrics.

[Yu09] Yu, Cheng-Bo; Qin, Hua-Feng; Cui, Yan-Zhe; Hu, Xiao-Qian: Finger-vein image recognition combining modified Hausdorff distance with minutiae feature matching. Interdisciplinary Sciences: Computational Life Sciences, 1(4):280–289, Dec 2009.

[ZS84] Zhang, T. Y.; Suen, C. Y.: A Fast Parallel Algorithm for Thinning Digital Patterns. Commun. ACM, 27(3):236–239, March 1984.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 231

Application of affine-based reconstruction to retinal point patterns

Mahshid Sadeghpour, Arathi Arakala, Stephen A. Davis, Kathy J. Horadam[1]

Abstract: Inverse biometrics that exploit the information of biometric references from comparison scores can compromise sensitive personal information of the users in biometric recognition systems. One inverse biometric method that has been very successful in regenerating face images applies an affine transformation to model the face recognition algorithm. This method is general and could apply to templates extracted from other biometric characteristics. This research proposes two formats to apply this method to spatial point patterns extracted from retina images and tests its performance on reconstructing such sparse templates. The results show that the quality of the reconstructed retina point pattern templates is lower than would be accepted by the system as mated.

Keywords: Retina, Biometric Template Protection, MDS-based Reconstruction, Inverse Biometrics, Affine-based Approximation.

1 Introduction

Irreversibility is one of the two critical requirements emphasised by ISO/IEC [IS11] to protect the *sensitive personal data* in biometric references [EC16]. Among reversibility attacks, those that exploit further information of data subjects, inverse biometrics, are classified into four groups based on the level of knowledge required by the attacker [GBG20]. The first group requires knowledge about the template format. The next group, hill-climbing attacks [Ad04, Ga13, Go14], requires access to the scores issued by the system. The third group requires access to the scores generated by the system as well as a set of imposter samples to be presented to the system. The last group are those attacks which require knowledge of the feature extraction method. The method in the third category, which will be reviewed and applied here, was proposed by Mohanty *et al.* [MSK06, MSK12]. Hereinafter we call it the *MSK* method.

The *MSK* algorithm models a biometric comparison algorithm using the scores issued by the system. An attacker needs to have access to a pool of imposter biometric samples (the *break-in set*) to present to the system. This attack is non-iterative. Having access to such a pool, the attacker can perform most of the attack offline without requiring iterative improvements of a presented sample.

The results in [MSK12] are very convincing in regenerating face images, and outperform hill-climbing attacks [Ad04]. So far, this attack has been tested on face recognition systems. The general consensus is that any biometric recognition system that issues the comparison scores is vulnerable toward the *MSK* attack. We would like to check the threat

[1] Mathematical Sciences, School of Science, RMIT University, Melbourne, mahshid.sadeghpour@rmit.edu.au, arathi.arakala@rmit.edu.au, stephen.davis@rmit.edu.au, kathy.horadam@rmit.edu.au

of this attack on different biometric characteristics, those with sparser representations in specific. To the best of our knowledge, it has not yet been tested on vascular biometric characteristics or on point pattern templates. However, the essence of the attack is general enough to be applied to fixed-length templates of other biometric characteristics.

This paper applies this attack to spatial point pattern templates (of varying lengths) extracted from retinal vascular graphs. It has been shown that retina can be accurately represented and compared using sparse templates comprising locations of feature points instead of the whole image [Ar16]. Our intuition was that this type of sparse biometric template might be more resistant to the *MSK* inverse attack.

We propose two formats for inputting the biometric information of the break-in sets as vectors of the same length. By its construction, the *MSK* algorithm should be able to reconstruct the break-in set very well. If the modelling does not do a good job in regenerating the break-in set, it probably would not successfully reconstruct target references. Our experiments show that neither format can reconstruct the break-in set well, nor can they successfully reconstruct the biometric references. The *MSK* algorithm will be reviewed briefly in section 2. The replication of results of Mohanty *et al.* [MSK06] will be presented in section 2.1. Section 3 gives two methods to apply this attack to retinal spatial point pattern templates and tests their performance reconstructing break-in sets and reference databases. Conclusions and future work appear in section 4.

2 Review of the *MSK* algorithm

The *MSK* algorithm approximates any comparison algorithm by an affine transform, given sufficiently many pairwise Euclidean distances between a pool of biometric samples (break-in set) $X = \{X_1, \ldots, X_K\}$. The pairwise distances between the break-in set samples are used to define an affine space where the representations of break-in set samples in the affine space, $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_K\}$, will have the same pairwise distances as in the Euclidean space. To find the approximating affine transformation $\mathbf{A}$ that maps the break-in set samples $X_i,\ 1 \leq i \leq K$, to the modeled vectors $\mathbf{Y}_i,\ 1 \leq i \leq K$, the *MSK* algorithm inputs the data of break-in templates $\mathbf{X}_i,\ 1 \leq i \leq K$, as columns of a matrix $\mathbf{X}$. Thus, the break-in templates are required to have the same length. For an unknown target sample X_t, if an attacker has access to the distance vector $d' = (d'_1, \ldots, d'_K)$ between X_t and the break-in samples, he can locate the transformed target $\mathbf{Y}_z$ in the affine space. Using a pseudo-inverse $\mathbf{A}^\dagger$ of affine approximation $\mathbf{A}$, the attacker can then reconstruct the target template $\mathbf{X}_z$ from its point $\mathbf{Y}_z$ in the affine space. Full details can be found in [MSK12].

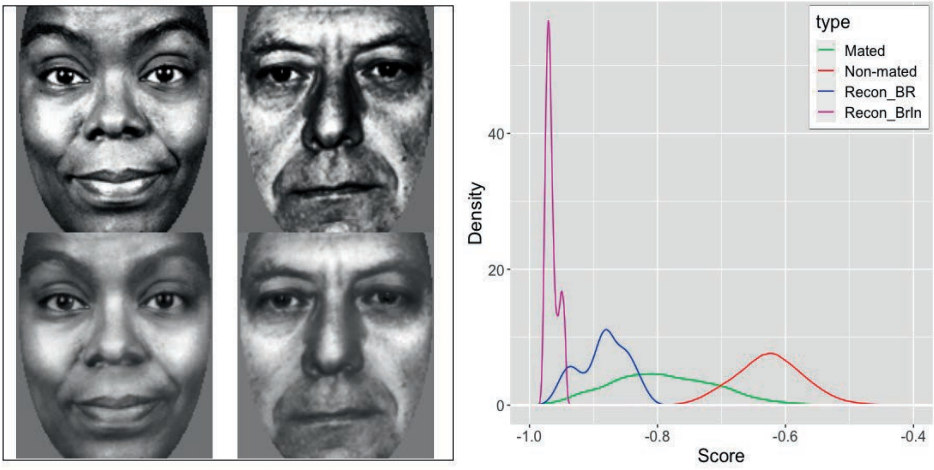
2.1 Reproducing the results of face image reconstruction

We converted the *MSK* Matlab code in [MSK06] to R code, then used it to model the PCA-based face recognition system [TP91] using FERET face database [Ph98, Ph00]. The underlying face recognition system applies Mahalanobis Cosine distance to compare the eigen-faces [Be03]. This experiment is performed to confirm that the R code is capable of

reproducing Mohanty *et al.*'s results in [MSK06]. The break-in set consists of 596 samples of 149 individuals from FERET. The images are 150×130 pixels, resulting in face image vectors of length $19,500$.

First, we tried to reconstruct the break-in set members. From an attacker's point of view, a successful modelling of the biometric algorithm would result in reconstructed break-in samples with high quality. The reason is that when a break-in sample is considered as the target, its distance vector to the break-in samples has 0 in one coordinate, as the target already exists in the break-in set. Figure 1a shows two of the original break-in samples from FERET on the top row and their corresponding reconstructed faces on the bottom row.

We then used the break-in set to reconstruct samples from 100 different target biometric references (BRs) in the FERET dataset and presented the reconstructed face images to the PCA-based system as probe samples.



(a) (b)

Fig. 1. a) Instances of some original faces from break-in set (top row) and their corresponding reconstructed faces (bottom row). b) Mahalanobis Cosine score distributions for mated (green), non-mated (red), and *MSK*-reconstructed samples (purple, reconstructed break-in set; and blue, reconstructed target BRs)

Figure 1b illustrates the distributions of Mahalanobis scores for mated, non-mated, and reconstructed samples in the PCA-based face recognition system. The distribution of scores achieved by comparing reconstructed reference samples to the original samples (the blue curve in Figure 1b) shows that all the reconstructed faces could be accepted by the system, which confirms the results in [MSK06]. The scores obtained by comparing the reconstructed break-in faces with the original break-in samples show that break-in faces are reconstructed with very high quality (purple curve in Figure 1b). Reconstructed break-in faces have better quality compared to reconstructed BRs.

3 Application of *MSK* algorithm to retina point patterns

This study endeavours to tune the *MSK* algorithm on spatial point pattern templates. The attacker has access to the comparison scores issued by the system, and is unable to gain access to the BRs. However, he has access to the break-in set. From the attacker's perspective, a smart method is to try and reconstruct the break-in set templates, first. Then, tune the attack using the knowledge obtained by reconstructing the break-in set. It is expected from the algorithm to successfully reconstruct each break-in template $\mathbf{X}_i,\ 1 \leq i \leq K$ since its corresponding data point $\mathbf{Y}_i,\ 1 \leq i \leq K$ exists in the modelled affine space (having zero distance to itself).

We conducted our experiments over a retinal vascular database, called ESRID (ECG Synchronised Retinal Image Database) [Ha12]. This database is collected by RMIT University and is accessible on request from the authors in [Ha12]. ESRID consists of 414 retinal images of 46 data subjects. Each individual in this dataset has 9 samples of their left eyes. The size of images in ESRID database is 2376×1584 pixels. The retinal point patterns are extracted from spatial graphs that are rescaled and centered on the optic disc. Rescaling sets the fovea on the point $(1,0)$ and the point patterns do not require further registration. The feature points from each image are extracted as real-valued spatial coordinates (x,y), and values can be negative. In experiments that reconstruct the break-in set, templates from the first data subject are considered as references, and the remaining 405 templates constitute the break-in set. In experiments that reconstruct BRs, we performed 46 experiments to reconstruct the BRs that are independent from the break-in set. Each of these 46 experiments reconstructs templates of one data subject using templates from every other data subject as break-in set.

We were interested in investigating the impact of the underlying comparison function on the performance of *MSK* reconstructing point pattern templates. We applied two different point pattern comparison functions: ICP (Iterative Closest Point) and MHD (Modified Hausdorff Distance) [DJ94] in our experiments.

3.1 Adaption of the spatial coordinates format

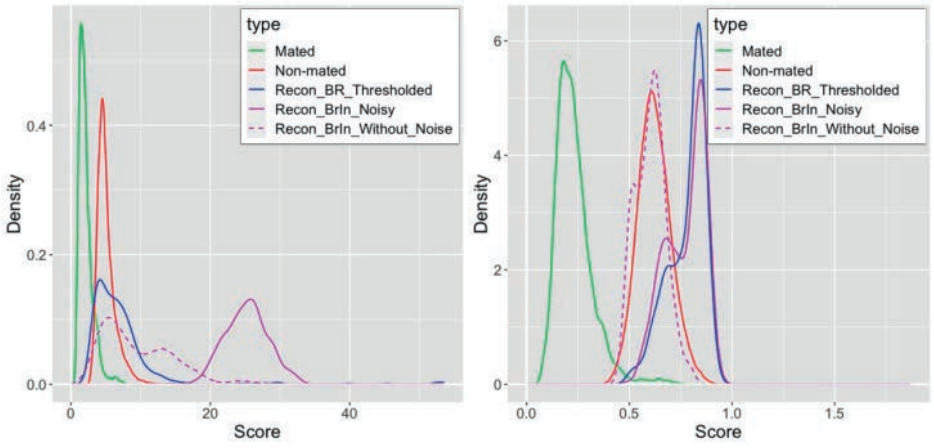
Here each $\mathbf{X}_i$ is a list of (x,y) coordinates of the points in the break-in sample X_i and would be read as a column vector with x-coordinates followed by y-coordinates. However, the sizes of $\mathbf{X}_i$s vary since different break-in samples have different numbers of points. The attacker needs to modify $\mathbf{X}_1,\ldots,\mathbf{X}_K$ to have the same dimensions by padding each $\mathbf{X}_i$ with enough $(0,0)$ points to increase its length to the maximum length template. After padding, each template will be of length $1,092$. Using *MSK*, any reconstructed templates will have length $1,092$ and a cluster of points close to $(0,0)$ caused by reconstructing the added $(0,0)$ coordinates.

To thoroughly study the impact of this introduced noise, we first focused on break-in samples, which an attacker should be able to reconstruct well (as they can for face images). The idea is that the attacker is well aware of the sizes of the break-in templates. Thus,

when $\mathbf{X}_z$ is generated for a break-in template, the attacker can truncate exactly the auxiliary points it contains. Figure 2 shows the scores obtained by comparing the reconstructed break-in spatial coordinates with the original break-in spatial coordinates both with noise (solid purple curve) and without noise (dashed purple curve) for MHD and ICP.

To estimate the effect of the added noise on reconstruction of BRs we checked the performance of the attack after discarding the points (x,y) from the reconstructed templates, where $-\theta \leq x,y \leq \theta$ for $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$. Since the performance of the attack using these values was not markedly different, we only present here the results for $\theta = 0.5$. The blue curves in Figure 2 represent the results of these experiments. Note that the plotting function in Figure 2a over-smooths the curve for reconstructed BRs. The smallest MHD score obtained by comparing the reconstructed BR templates and their corresponding targeted templates equals 2.782. This score is greater than the smallest score for non-mated comparisons which is 2.408.

In Figures 2, the distribution curves for scores obtained by comparing the reconstructed templates with the original templates illustrate that when treating retinal point patterns as spatial coordinates, the reconstructed templates would not be accepted by the system as mated templates. In fact, even discarding all noise introduced to the break-in templates (the dashed purple curves) does not improve the performance of the attack enough to reconstruct templates that could be considered by the comparator as mated templates.



(a) (b)

Fig. 2. Score distributions for mated (green), non-mated (red), and *MSK*-reconstructed templates (blue, with ESRID BRs; solid purple, with ESRID break-in set with noise; dashed purple, with ESRID break-in set noise-reduced) using MHD (a) and ICP (b).

3.2 Adaption of the binary image format

Here, we treat the templates as binary images with discretised intensity values equal to either 255, if a pixel contains a feature point, or 0, otherwise. Feature points in these tem-

plates are real-valued spatial coordinates (x,y). To identify the pixels that contain feature points, the original spatial coordinates were shifted, scaled and rounded to ensure that all the coordinates were positive integers. To minimise rounding error, we applied the inverse transformation to the rounded coordinates and calculated the MHD and ICP distance between the inverted values and the original (x,y) coordinates, with scaling values $1,5,10$ and 15. For accuracy and speed, we selected $s = 10$, resulting in templates of length $1,254,400$.

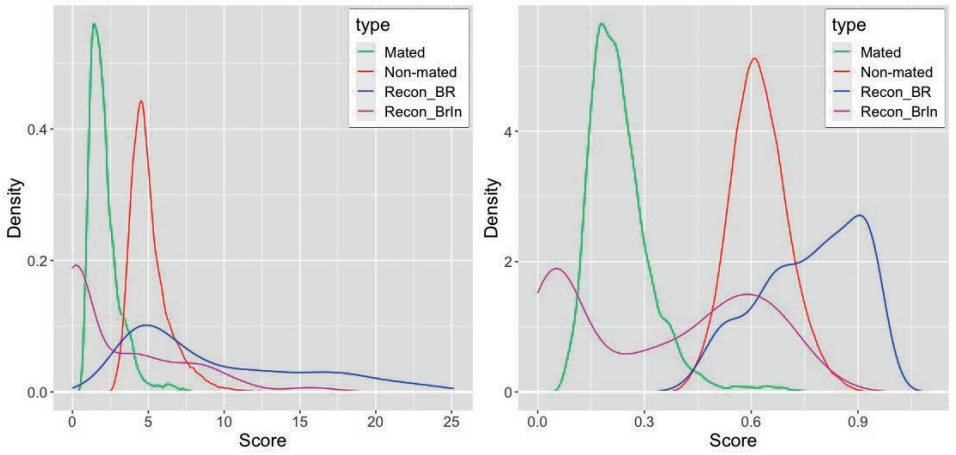
After performing the attack on a target sample X_t, the returned intensity values in the vector $\mathbf{X}_z$ are greyscale, and must be binarised by thresholding. An attacker can find threshold(s) θ for the reconstructed break-in set then use this information to binarise reconstructed BRs. Following this approach we found that each greyscale break-in reconstruction had its own threshold. If $\theta = 128$, no binarised break-in template is reconstructed. Optimising over the thresholds we found for each break-in reconstruction gave $\theta = 25$ and $\theta = 38$, respectively, for MHD-based and ICP-based systems. However, not all binarised break-in reconstructions with these uniform thresholds will contain feature points. The purple curves in Figure 3 illustrate the performance of the reconstructed break-in set using the optimum thresholds.

Then, we applied those values as thresholds to reconstruct BRs. With $\theta = 25$ (resp. $\theta = 38$) for the MHD-based (resp. ICP-based) system, 32.6% (resp. 28.51%) of the binarised reconstructed BR templates had no feature points in them. We discarded their scores when plotting performance. The blue curves in Figure 3 illustrate the performance of reconstructing BRs using the optimum thresholds.

We see that when treating retinal point patterns as binarised images, the reconstructed BR templates would not be accepted by the system as mated templates. In fact, many of the break-in templates cannot be reconstructed well enough to be considered by the comparator as mated.

4 Conclusion

We applied the *MSK* algorithm to two formats of retina point pattern templates in this paper. The experimental results showed that the performance of this attack on retinal vascular point patterns is not comparable with that on face images. From our point of view, the approximation procedures in calculating $\mathbf{Y}_i$ s and $\mathbf{A}^{\dagger}$ skew the the values more than a point pattern comparison algorithm can tolerate. Image-based face comparison algorithms can tolerate this amount of deviation since there is a high amount of continuity among their values, whereas for point patterns this is not the case. The performance of this attack on retinal vascular point patterns is so poor that it might not be considered as a threat to privacy and security of the users and their templates. The results using our proposed formats showed that the attack is not even successful in reconstructing break-in templates. For our future work, we would like to check the performance of the *MSK* attack on the point patterns extracted from other vascular biometric characteristics that have reasonable recognition accuracy. We are also interested in exploring the reconstruction of point pat-



(a) (b)

Fig. 3. Score distributions for mated (green), non-mated (red), and *MSK*-reconstructed templates (blue, with ESRID BRs; purple, with ESRID break-in set) using MHD (a) and ICP (b).

tern and graph-based vascular templates using a hybrid reconstruction method that applies the affine-based reconstruction algorithm followed by a hill-climbing attack.

Acknowledgements The first author was supported by an RMIT University RD Gibson Grant. We thank the reviewers for their comments, the authors in [MSK06] for providing us with their codes for reconstructing faces, and NCI Australia for computational resources. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

[Ad04] Adler, Andy: Images can be regenerated from quantized biometric match score data. In: Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No. 04CH37513). volume 1. IEEE, pp. 469–472, 2004.

[Ar16] Arakala, Arathi; Davis, Stephen A; Hao, Hao; Horadam, Kathy J: Value of graph topology in vascular biometrics. IET Biometrics, 6(2):117–125, 2016.

[Be03] Beveridge, Ross; Bolme, David; Teixeira, Marcio; Draper, Bruce: The CSU face identification evaluation system userâs guide: version 5.0. Computer Science Department, Colorado State University, 2(3):1–29, 2003.

[DJ94] Dubuisson, M-P; Jain, Anil K: A modified Hausdorff distance for object matching. In: Proceedings of 12th international conference on pattern recognition. volume 1. IEEE, pp. 566–568, 1994.

[EC16] European Council, General Data Protection: Regulation of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. Official Journal of the European Union (OJ), 59(1-88):294, 2016.

[Ga13] Galbally, Javier; Ross, Arun; Gomez-Barrero, Marta; Fierrez, Julian; Ortega-Garcia, Javier: Iris image reconstruction from binary templates: An efficient probabilistic approach based on genetic algorithms. Computer Vision and Image Understanding, 117(10):1512–1525, 2013.

[GBG20] Gomez-Barrero, Marta; Galbally, Javier: Reversing the irreversible: A survey on inverse biometrics. Computers & Security, 90:101700, 2020.

[Go14] Gomez-Barrero, Marta; Galbally, Javier; Morales, Aythami; Ferrer, Miguel A; Fierrez, Julian; Ortega-Garcia, Javier: A novel hand reconstruction approach and its application to vulnerability assessment. Information Sciences, 268:103–121, 2014.

[Ha12] Hao, Hao; Sasongko, Muhammad B; Wong, Tien Y; Azemin, Mohd Zulfaezal Che; Aliahmad, Behzad; Hodgson, Lauren; Kawasaki, Ryo; Cheung, Carol Y; Wang, Jie Jin; Kumar, Dinesh K: Does retinal vascular geometry vary with cardiac cycle? Investigative ophthalmology & visual science, 53(9):5799–5805, 2012.

[IS11] ISO/IEC JTC1 SC27 IT Security Techniques: International Standard on Biometric Information Protection. Standard ISO/IEC24745, International Organization for Standardization, 2011.

[MSK06] Mohanty, Pranab; Sarkar, Sudeep; Kasturi, Rangachar: Privacy & security issues related to match scores. In: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06). IEEE, pp. 162–162, 2006.

[MSK12] Mohanty, Pranab; Sarkar, Sudeep; Kasturi, Rangachar: , Reconstruction of biometric image templates using match scores, April 24 2012. US Patent 8,165,352.

[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.

[Ph00] Phillips, P Jonathon; Moon, Hyeonjoon; Rizvi, Syed A; Rauss, Patrick J: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on pattern analysis and machine intelligence, 22(10):1090–1104, 2000.

[TP91] Turk, Matthew; Pentland, Alex: Face recognition using eigenfaces. In: Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition. pp. 586–587, 1991.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 239

Facial Profiles Recognition Using Comparative Facial Soft Biometrics

Malak Alamri[1], Sasan Mahmoodi[2]

Abstract: This study extends previous advances in soft biometrics and describes to what extent soft biometrics can be used for facial profile recognition. The purpose of this research is to explore human recognition based on facial profiles in a comparative setting based on soft biometrics. Moreover, in this work, we describe and use a ranking system to determine the recognition rate. The Elo rating system is employed to rank subjects by using their face profiles in a comparative setting. The crucial features responsible for providing useful information describing facial profiles have been identified by using relative methods. Experiments based on a subset of the XM2VTSDB database demonstrate a 96% for recognition rate using 33 features over 50 subjects.

Keywords: Comparative Soft Biometrics, Profile Face Recognition, Profile Facial Attributes, Ranking.

1 Introduction

Due to increasing security threats around the world, there is an urgent need for more advanced technologies in the field of biometrics, particularly in facial recognition. Popular authentication methods in diverse security systems involve identity verification based on the identity card of an individual and identification based on biometric measurements. In their 2018 survey, Abdelwhab et al. explain that unlike other methods, biometrics are unique for each individual; cannot be easily transferred; are readily available; and cannot be easily borrowed, forgotten, shared, stolen, or observed [AV18].

The soft biometric provides extra knowledge for higher recognition by using comparative description based on eyewitness testimonies from a scene of a crime. Soft biometrics are dynamic features of an individual such as age, skin colour, height, ethnicity, and face dimensions, which provide additional information provided by eyewitnesses to improve the accuracy and reliability of traditional biometrics or to perform recognition for cases where there is no recoding system in the scene of the crime and there are only eyewitness testimonies to describe the criminal. Although traditional biometrics play a primary role in recognition and detection, recent research shows that the use of soft biometrics has significant potential in different applications such as identification, and identity verification. For instance, Klare et al. utilized hand-drawn sketches and compared them with facial components using two experimental methods for identifying suspects in criminal investigations

[1] Physical Sciences and Engineering, Electronics and Computer Science, University of Southampton, Highfield, M.Alamri@soton.ac.uk
[2] Physical Sciences and Engineering, Electronics and Computer Science, University of Southampton, Highfield, S.Mahmoodi@soton.ac.uk

[Kl14]. The approach was relatively successful for enhancing the recognition/identification of individuals through verbal description, particularly by victims of criminal activities.

Various studies have assessed the efficacy of profile face images. For instance, in [YEE19], authors show how the profile face images can affect the accuracies and they found that age and gender classification can achieve high accuracies by combining ear and profile face images which contain a valuable information. Moreover, [BU17] and [ZW11] demonstrate that estimating ages from ear and side-view of face images leads to a promising performance in recognition rate.

Although there are many research studies on facial profile attribute analysis, only a few are concerned with the analysis of facial profile attributes for biometric purposes. Facial recognition remains significantly affected by the wide variations of pose. The pose problem makes the training of face retrieval algorithms challenging. In fact, effective recognition requires the capture of numerous face images at different angles for the same person. The existing systems in the literature do not provide a large volume of annotated side view faces. Our main contribution in this work is to propose facial profiles as a viable biometric system in a soft biometric framework. In summary, the contributions of our work are listed in more details as:

- It establishes a soft biometric system with face profiles to highlight the significance of profile (or side view) in biometric recognition.
- It proposes a new set of semantic profile facial attributes along with their comparative labels.
- It identifies the important attributes that enable efficient recognition of an individual using the profile face.

The remainder of the paper is organized as follows: Section 2 explains the research approaches and the use of a ranking system with semantic attributes and labels. Section 3 is a description of the experimental platform and discusses the results. Finally, Section 4 draws some conclusions which outlines our research and discusses the future work.

2 Methodology

2.1 Attributes Definition

In this paper, we analyse profile face attributes based on approaches used in previous studies. In our method a new set of facial profile attributes are proposed for comparative soft biometrics for recognition and identification. We also use some existing soft biometric features previously proposed in, [ANH16a], [ANH16b], [ANH17], which describe the important traits of a human face, e.g. shape of an eyebrow, eye and nose and allow the definition of 26 attributes relevant for extracting the identity of each face. Our proposed new attributes are nostril size, nose tip, face profile height, face profile width, ear-to-head

ratio, ear-to-nose distance and ear-to-chin distance, because they can, intuitively, describe or be associated with a facial profile.

2.2 Profile Facial Dataset

We used the XM2VTSDB dataset for this research. This dataset is established and maintained by the University of Surrey. The multimodal database hosts numerous speech recordings, video sequences, and facial images from 295 subjects [Me99]. The recordings of images in XM2VTSDB spanned an extended period, involving four sessions, to allow significant variation in the appearance of the subjects. For instance, the recordings preserved in the resources are likely to have individuals with variances in shape, facial hair, and hairstyles. Fig.1 shows example images in XM2VTSDB face profile dataset.

Fig. 1: Example images in facial profiles in the sample XM2VTSDB dataset

2.3 Relative Rating of Attributes from Comparative Labels

In this paper, we have used a comparative scheme to rank subjects based on their attributes. These comparative labels allow systems and individuals to compare relative features among various subjects to avoid data biases and human (labelers) errors in comparison with a categorical framework. Consequently, the soft biometrics features are generated based on this comparative scheme for our soft biometric system [ANH17].

This study utilized a 4-point bipolar scale for the comparative labels associated with attributes (see Tab. 1). The label values are 1 for ”More A,” 0 for “Same,” -1 for More B/Less A,” and -2 for “Cannot see”.

2.4 Data Acquisition Through Crowdsourcing

Effective labelling of a dataset is of critical importance to the research process in a soft biometric framework. As a result, a significant portion of the existing literature on soft biometrics utilizes a crowdsourcing platform for labelling the datasets. Such an approach provides a reliable method for the analysis of traits and labels [ANH16a], [ANH17]. Thus, the construction and monitoring of crowdsourcing annotations involves the use of the Appen platform in the collection of labels. This platform guarantees high-quality annotations by spreading analyses and encouraging customers to use a range of answers. It also identifies and rejects dishonest responses. A total of 50 subjects with four profile samples obtained from the XM2VTSDB dataset are used for this experiment. Tab. 2 presents an overview on the crowdsourcing of comparative labels.

No.	Soft Traits	Comparative Labels			
		1	0	-1	-2
1	Eyebrow length	More Long	Same	More Short	Cannot see
2	Eyebrow shape	More Raised	Same	More Low	Cannot see
3	Eyebrow thickness	More Thick	Same	More Thin	Cannot see
4	Spectacles	More Covered	Same	Less Covered	Cannot see
5	Eye-to-eyebrow distance	More Large	Same	More Small	Cannot see
6	Eye lashes	More Long	Same	More Short	Cannot see
7	Eye size	More Large	Same	More Small	Cannot see
8	Nose-to-mouth distance	More Long	Same	More Short	Cannot see
9	Nostril size	More Wide	Same	More Narrow	Cannot see
10	Nose tip	More Pointed Down	Same	Less Pointed Down	Cannot see
11	Nose size	More Large	Same	More Small	Cannot see
12	Lips thickness	More Thick	Same	More Thin	Cannot see
13	Face profile height	More Long	Same	More Short	Cannot see
14	Face profile width	More Wide	Same	More Narrow	Cannot see
15	Skin smoothness	More Smooth	Same	Less Smooth	Cannot see
16	Skin condition	More Clear	Same	More Pimples	Cannot see
17	Forehead hair	More Forehead Hair	Same	Less Forehead Hair	Cannot see
18	Ear size	More Large	Same	More Small	Cannot see
19	Ear orientation with respect to head	More Further from head	Same	More Close to head	Cannot see
20	Ear-to-head ratio	More Large	Same	More Small	Cannot see
21	Ear-to-chin distance	More Further	Same	More Close	Cannot see
22	Ear-to-nose distance	More Large	Same	More Small	Cannot see
23	Cheek shape	More Flat	Same	More Prominent	Cannot see
24	Cheek size	More Large	Same	More Small	Cannot see
25	Chin and jaw shape	More Receding	Same	More Protruding	Cannot see
26	Double chin	More Large	Same	More Small	Cannot see
27	Chin height	More Large	Same	More Small	Cannot see
28	Neck length	More Long	Same	More Short	Cannot see
29	Neck thickness	More Thick	Same	More Thin	Cannot see
30	Age	More Old	Same	More Young	Cannot see
31	Gender	More Masculine	Same	More Feminine	Cannot see
32	Skin colour	More Dark	Same	More Light	Cannot see
33	Figure (shape)	More Fat	Same	More Thin	Cannot see

Tab. 1: Soft profile face biometric attributes and comparative labels

2.5 Ranking by Relative Profile Face Attributes

The Elo rating system is a popular algorithm for ranking players in chess. The system ranks players by using variances between the actual results in a game and expectations. The effectiveness of the scale is making it popular in other fields, such as soft biometrics recognition [RNS13]. The biometric signatures which are feature vectors composed of the relative strength of attributes based on comparative labels, will be generated by Elo rating system. Almudhahka et al. uses the Elo system in their study to evaluate the comparative rates between features from biometric signatures and comparative labels [ANH17].

Total number of labelers per question	15
Total traits comparison per subject	3.960
Total number of images	200
Average number of comparison per subject	2
Total trusted judgment	198.000

Tab. 2: The statistics of crowd-sourcing task for XM2VTSDB dataset

The use of comparative soft biometrics involves distinct processes and activities. The systematic process begins with the construction of a dataset based on the Appen platform. The next step is the conversion of comparisons made by labelers into ranks using Elo rating system. Such ranks then provide a set of feature vectors for profile faces for each image. Finally, the k-NN classifier is used to calculate the recognition rate.

3 Experiments

3.1 Correlation Analysis

Pearson's correlation r, helps spot linear dependencies between the attributes. Equation (1) shows how Pearson's correlation r between variables x and y is measured [To15]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where x and y are independent variables, used to describe traits on the facial profiles, in this regard, x_i and y_i are two distinct labels, representing i^{th} annotation of a given subject.

The collected profile face comparisons in the study have significant correlations, as illustrated in Fig. 2 Dark brown color on the cells represents traits with high positive correlation; and dark green color corresponds to a strong negative correlation; and white/light cells show the absence of a linear correlation. A positive or negative correlation between features and labels expresses dependencies between two traits. Features with negative relationships are highly reliable for distinguishing individuals from others. Fig. 2 shows a positive correlation between age and profile face width. Gender and eyebrow thicknesses also have a positive correlation. Moreover, there is a significant relationship between ear size and age. However, age and eyebrow length have a negative correlation.

3.2 Discriminative Power of Facial Profiles

The improvement of efficiency and accuracy requires a reduction in the number of non-useful features. In this study, feature analysis and orderings facilitated feature set selection through mutual information (MI) and sequential floating forward selection (SFFS)[SSZ13].

Mutual Information (MI)

Mutual information (MI) measurement is used to assess the significance of the contribution of a profile attribute and MI is defined as:

$$MI = I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \ln \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2}$$

Where X is a variable representing an attribute, and Y is an attribute representing a label. The computation of MI using the two traits with p(x, y) is the probability density function for X and Y. In this regard, p(x) and p(y) are marginal probability density functions.

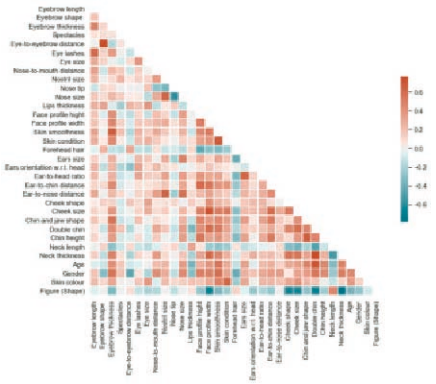


Fig. 2: The correlation matrix between the profile face attributes

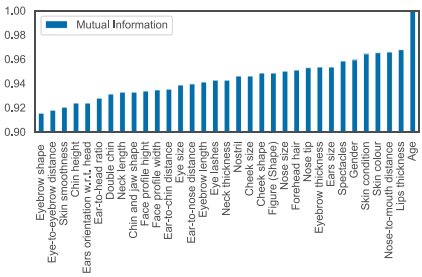


Fig. 3: Normalized MI for each of the 33 attributes with the target variable (label)

MI is used to evaluate the discriminative power of the 33 profile attributes by computing its value for each attribute jointly with the label. Fig. 3 illustrates that skin colour, nose-to-mouth distance, lips thickness, and age have high scores compared to eyebrow shape and eye-to-eyebrow distance, which have low MI scores.

Sequential Floating Forward Selection (SFFS) Algorithm

The SFFS algorithm helps rapidly identify the best set of features which contribute in the recognition. The SFFS algorithm begins with an empty set (d=0). For each iteration, it tries to find the best attribute that can be added to improve the recognition rate. Then, in the same iteration, it tries to remove a feature that was selected in past iterations in order to further improve the recognition. The algorithm repeats the steps in each iteration until no improvement can be realized. In our experiments, SFFS identified 17 attributes among the 33 features (3,4,7,9,11,12,13,15,16,17,18,24,29,30,31,32,33) (see Tab. 1). The features selected by this algorithm has led to a final value of 98% accuracy for recognition rate.

3.3 Recognition Performance of Facial Profiles

To measure the recognition rate, a Leave-One-Out Cross Validation (LOOCV) strategy is employed in this paper. We have employed k-NN as a basic classifier. In this study, by

using all the attributes proposed in Tab. 1, an accuracy of 96% for the recognition rate is achieved, as illustrated in Fig. 4. The dataset in this study contained 50 subjects with four images per subject. There are therefore 200 face profile images in our dataset. As a result, there are almost 1000 questions per image. One image in the dataset was used as a testing image and the remaining ones as the training set. The k-NN classifier used a training split described by 1-vs-rest.

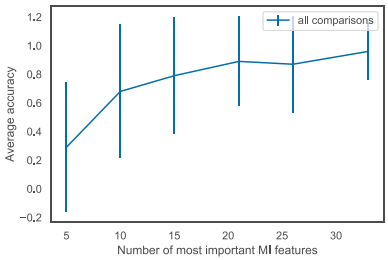


Fig. 4: The results of the accuracy with (≈ 1000) comparisons per subject

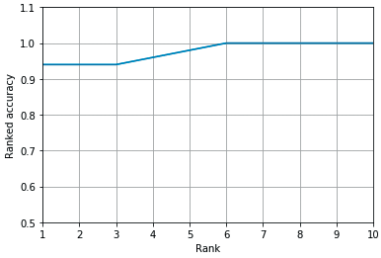


Fig. 5: Recognition via CMC performance of this study for the 33 attributes and 25% test samples

In our study, we have achieved 96% recognition rate with 50 subjects. This is comparable to Nawaf et al. study where 100 subjects with front view faces were considered and 100% recognition rate was achieved [ANH16a]. Obviously a lower recognition rate is expected with face profiles (side view faces), since less information is available to labelers with face profiles. However our recognition rate of 96% indicates face profiles carry enough information to be considered as an important biometric modality. In fact, since we use Elo rating system, it seems that the subjects preserve their ranks independently of the view point. Future works will address the learning of a latent feature space, that is adapted for view independent face recognition, based on soft biometrics.

The CMC curve is a crucial tool for assessing retrieval performance [DR13]. The metric encapsulates recognition accuracy by employing the k-NN method. Fig. 5 depicts the recognition performance by facial profile traits using soft facial traits. In this curve, the first candidate has 96% accuracy, which increases with the improvement of the number of candidates to 100% at rank-6.

4 Conclusions and Future Work

This paper proposes a novel biometric system based on facial profiles in a soft biometric framework. The study proposes and evaluates a list of semantic human facial profile attributes, and it also introduces comparative labels to facilitate the assessment of comparative soft biometrics. Our numerical analysis in this paper demonstrates that face profiles can be considered as an important biometric modality.

Future work will focus on increasing the number of subjects in XM2VTSDB dataset. We also plan to find corresponding features by using computer vision techniques in a tradi-

tional biometric framework to allow profile face identification to show that the traits proposed in this paper are important in both the soft biometric and the traditional biometric.

References

[ANH16a] Almudhahka, N; Nixon, M; Hare, J: Human face identification via comparative soft biometrics. In: 2016 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). IEEE, pp. 1–6, 2016.

[ANH16b] Almudhahka, N Y; Nixon, M S; Hare, J S: Unconstrained human identification using comparative facial soft biometrics. In: 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–6, 2016.

[ANH17] Almudhahka, N Y; Nixon, M S; Hare, J S: Semantic face signatures: Recognizing and retrieving faces by verbal descriptions. IEEE Transactions on Information Forensics and Security, 13(3):706–716, 2017.

[AV18] Abdelwhab, A; Viriri, S: A survey on soft biometrics for human identification. Machine Learning and Biometrics, p. 37, 2018.

[BU17] Bukar, A M; Ugail, H: Automatic age estimation from facial profile view. IET Computer Vision, 11(8):650–655, 2017.

[DR13] DeCann, B; Ross, A: Relating roc and cmc curves via the biometric menagerie. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, pp. 1–8, 2013.

[Kl14] Klare, B F; Klum, S; Klontz, J C; Taborsky, E; Akgul, T; Jain, A K: Suspect identification based on descriptive facial attributes. In: IEEE International Joint Conference on Biometrics. IEEE, pp. 1–8, 2014.

[Me99] Messer, K; Matas, J; Kittler, J; Luettin, J; Maitre, G: XM2VTSDB: The extended M2VTS database. In: Second international conference on audio and video-based biometric person authentication. volume 964, pp. 965–966, 1999.

[RNS13] Reid, D A; Nixon, M S; Stevenage, S V: Soft biometrics; human identification using comparative descriptions. IEEE Transactions on pattern analysis and machine intelligence, 36(6):1216–1228, 2013.

[SSZ13] Shirbani, F; Soltanian Zadeh, H: Fast SFFS-Based Algorithm for Feature Selection in Biomedical Datasets. AUT Journal of Electrical Engineering, 45(2):43–56, 2013.

[To15] Tome, P; Vera-Rodriguez, R; Fierrez, J; Ortega-Garcia, J: Facial soft biometric features for forensic face recognition. Forensic science international, 257:271–284, 2015.

[YEE19] Yaman, D; Eyiokur, F I; Ekenel, H K: Multimodal Age and Gender Classification Using Ear and Profile Face Images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0, 2019.

[ZW11] Zhang, G; Wang, Y: Hierarchical and discriminative bag of features for face profile and ear based gender classification. In: 2011 International joint conference on biometrics (IJCB). IEEE, pp. 1–8, 2011.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 247

Development and empirical optimization of an electrochemical analysis cell for the visualization of latent fingerprints and their chemical adhesives

Tommy Bergmann[1],Sebastian Gottschall[2],Enrico Fuchs[2],Oliver Berlipp[1], Dirk Labudde[1,3]

Abstract:

Fingerprint analysis played a major role in the investigation of criminal offences for the past 100 years and is often the sole means of criminal identification [YA04]. Electrochemical analysis can yield important additional evidence like fingerprint age, biological age and gender of its creator as well as chemical adhesives [GRW12]. Additional gained characteristics through electrochemical analysis can supplement latent or incomplete fingerprints. In previous work a ruthenium-complex based solution was used as illuminant. Since luminol is readily available and is used in many forensic applications, the presented paper will focus on luminol as an alternative chemical for the ECL-aided visualization of fingerprints. Experiments were conducted by creating an electrochemical reaction inside a purpose build analysis cell. Eccrine, sebaceous glandlike and vaseline contaminated fingerprints were created on a stainless-steel plate placed inside the cell and investigated while applying direct current. Aim of this research was to investigate which kind of fingerprints can be visualized and which quality of the resulting images can be reached using luminol as illuminant. The used laboratory power supply created a strong light reaction at the start of each experiment revealing potential for further enhancement of the image quality. Eccrine dactyloscopic evidence showed no visible results. For sebaceous glandlike fingerprints age was discovered to significantly influence image quality.

Keywords: latent fingerprints (LFP), electrochemoluminescence (ECL), luminol, chemical adhesives (substances), gender determination, age determination, information of fingerprints, forensic science.

1 Introduction

1.1 Background

The use of forensic dactyloscopy for suspect indentification is as old as criminalistic itself. References to this can be seen in ”System und Praxis der Daktyloskopie” from *Heindl*. Also, *Heindl* clearly describes the historical development of dactyloscopy, which has already undergone several innovations in the course of its development [He22]. Today, fingerprints are even considered more valuable evidence than deoxyribonucleic acid (DNA)

[1] University of Applied Sciences, Computer-and Biosciences, Technikumpl. 17, 09648 Mittweida, Deutschland, Email: {*firstname.name*}@hs-mittweida.de
[2] Fraunhofer-Institut für Verfahrenstechnik und Verpackung IVV, Institutsteil Verarbeitungstechnik, Heidelberger Straße 20, 01189 Dresden, Deutschland, Email: {*firstname.name*}@ivv-dd.fraunhofer.de
[3] Fraunhofer-Institut für Sichere Informationstechnology SIT, Rheinstrasse 75, 64295 Darmstadt, Deutschland

[HS07]. A new approach for the visualization of latent fingerprints is the use of electrochemical luminescence (ECL) reactions. In ECL reactions the luminescence is generated electrochemically by applying an electrical potential e.g. to a luminol, ruthenium, or rubrene solution. The resulting intermediates are subjected to an immense exergonic reaction in order to reach an energetically higher state. In the further course of the process, the relaxation leads to a transition to the energetically lower state, whereby the energy difference can be observed in the form of light. ECL reactions are already proven in analytical applications because they are highly sensitive and can be used selectively by applying a potential [GA13, FBK09, Va16, PS74]. For example, *Beresford et al.* describe visualization by spatially selective deposition of an electrochromic polymer (polyaniline). The electrochemical process is inhibited by the fingerprint and a negative image is created. The advantages of their method is an increase in contrast by varying the applied electric potential. Also, the electrochromic coating results in a longevity of the evidence [BH10]. In the work of *Jasuja et al.* an aqueous electrolyte solution was used, which made it possible to visualize latent fingerprints on deformed surfaces (aluminum foil) [Ja15]. Additional examples are provided in the review‘s of *Su et al.* [Su16] and *Yamashita et al.* [YF11].

The work of *Xu et al.* shows that the combination of electrochemistry and forensic dactyloscopy has a considerable advantage. For example, explosive residues can be detected [Ad11, LZJ06]. Due to the difference in brightness on the electrode caused by this reaction and the fingerprint residue lying thereon and blocking the electron exchange a contrast is generated which results in high-resolution images of the fingerprint.

[a] [b]

Fig. 1: [a] Experimental setup with recommended chemicals published by *Xu et al.*; Reproduced by our research group [b] high-contrast and high-resolution images of fingerprints based on the method described by *Xu et al.*

Furthermore, electrochemical scanning methods are used to detect spatial differences in the electrochemical reactivity and surface exit work of fingerprints. It is of special interest to reach a resolution which makes pores visible as demonstrated by *Xu et al.* [Xu12]. We could reproduce this high quality images using their recommended chemicals (tris(2,2’-bipyridyl) ruthenium(II) ([Ru(bpy)3]2+) and tri-n-propylamine (TPrA) as illustrated in Fig. 1. All minutiae (islands, inclusions, branches, bridges, etc.), even sweat pores within the papillary ridges, are clearly visible. However the high price and poor availability of the used ruthenium-solution leads to the demand for alternatives. The use of luminol to make blood evidence visible with the help of a catalyst has been a common practice since 1937 [Sp37]. It is also used in immunoassays as a part of an antibody reaction [Ji13]. Due to

its wide usage in forensics using luminol for ECL-reactions is a cost efficient and obvious approach. Therefore the presented research will focus on the development of luminol as an alternative chemical for the ECL-aided visualization of fingerprints.

2 Materials and Methods

Development of an electrochemical analysis cell An analysis cell with a two electrode system was constructed out of available stainless steel components, a petri dish and completed by a purpose built part 3 D printed from polyethylene terephthalate (PETG). A plate electrically connected through a screw forming the base electrode was placed at the bottom of the petri dish. The insulation distance of 1 mm between both electrodes was realized with the 3 D printed part enclosing the second electrode as well as restricting the cells active area to a circular diameter of 28 mm while reducing the necessary liquid volume to 2.4 ml. Fig. 2 illustrates the construction of the cell.

Fig. 2: [a] Overview of the components, [b] construction sketch cross-section and [c] structured analysis cell. The construction as seen in the picture consists of a petri-Dish, they contain two electrodes (stainless steel -plate, -nut), separated by a plastic insulator.

Both electrodes were connected using crocodile clip equipped wires. While the wire to the stainless steel nut was constantly connected to a laboratory power supply the second electrode was switched by plugging and unplugging it. The influence of ambient light fluctuation was eliminated by conducting all experiments in a darkroom and using a camera for observation. Additionally, two ultra violet (UV) lamps [see Tab.1] illuminating the fingerprints at 45 degree angle were positioned left and right of the camera.

Materials	Manufacturer
Laboratory power supply unit: ”PPS-11360”	VOLTCRAFT
Fine balance: ”New Classic MF”	METTLER TOLEDO
Camera: ”VCXU-51C”	BAUMER
Lens: TV ZOOM Lens S6x11 11.5-69 mm	SPACECOM
Drying cabinet ”VC 0020”	Vötsch Industrietechnik
UV lamp: Synergy 21 LED Prometheus UV V2	ALLNET GmbH

Tab. 1: laboratory equipment

Luminol Solution In the experiments undiluted (0.025 mol/ l) and diluted luminol solution was tested. For the preparation of the solution, 0.44 g luminol was dissolved in 3 ml hydrogen peroxide (NaOH) with a purity of 50 % [Ea11]. Subsequently, 97 ml deionized water was added. For the further experiments, this luminol solution was used as the basis for the undiluted version or, with the addition of 100 ml deionized water, for the diluted version.

Preparation of the fingerprints For the following experiments the fingerprints (thumb and index finger) of a single person were used. Before the application of the fingerprints, the person was instructed to wash their hands with soap and then rinse them with lukewarm water. Drying was done by air. For the transfer of eccrine fingerprints, powder-free rubber gloves were worn for 10 minutes to stimulate sweat production. To get sebum with fingerprints, the person touched the forehead, the lateral nostrils, and the areas behind the ears with their fingers. Finally, to produce vaseline-containing fingerprints, contact was made with commercially available vaseline, which was wiped off on external surfaces before the fingerprint was transferred. All images of fingerprints listed in chapter 3 were transferred to a stainless steel plate, which then was included in the electrochemical analysis cell. For the transfer, the contact time was about one minute.

Fingerprint visualization The analysis cell was aligned so that the fingerprint was in the centre of the stainless steel nut and a reference picture was taken under UV exposure. Then 5 ml of the luminol solution was added to the fingerprint. Furthermore 0.25 ml of the hydrogen peroxide solution was added and a current of 2 amperes (A) was applied and a picture was taken. The voltage was between 8 V and 10 V in all experiments. This resulted in emission of light that occured everywhere in the solution, except at the adhesion (fingerprint) itself.

3 Results

In the experiments light emission could be observed everywhere in the solution except for the adhesion or the sebaceous fingerprint. The intensity of the emitted light shortly peaked at the start of each experiment when the power supply was connected. The ECL reaction resulted in a useful contrast only when the fingerprint was placed onto the anode.
Fig. 3 shows a comparison of the visibility of the characteristics of differently aged sebaceous fingerprints. In Fig. 3 [a] a partial impression of the fresh fingerprint with characteristic values and optical anomaly (scar) is shown. In Fig 3 [b] a 16 h aged fingerprint is visible, including more detailed anatomic features. The Comparison of the visibility of the anatomic features of the fresh and the aged sebaceous fingerprint revealed that the aged fingerprint created better optical results. The picture in Fig. 3 [a] was taken several seconds delayed to the application of the electrical current resulting in a vivid reaction growing form the outside inwards.

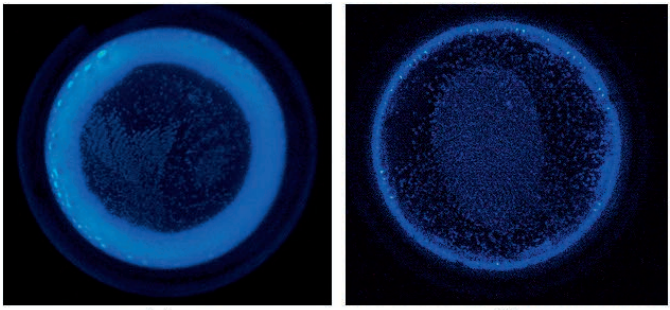


Fig. 3: Images created with ECL of two sebaceous fingerprints [a] fresh and [b] aged

As with *Xu et al.*, it could be experimentally confirmed in Fig. 4 [a], [b] that vaseline adhesion can also block signals. Details of the anatomic features were not visible, only the outline of the fingerprint. All areas of the fingerprint coated with vaseline showed a specific reaction (a bubble-like pattern). Fig. 4 demonstrates the results of fingerprints covered with vaseline. The experiment was carried out as described in chapter 2. The undiluted luminol solution [see 2 Luminol Solution] was used for this purpose.

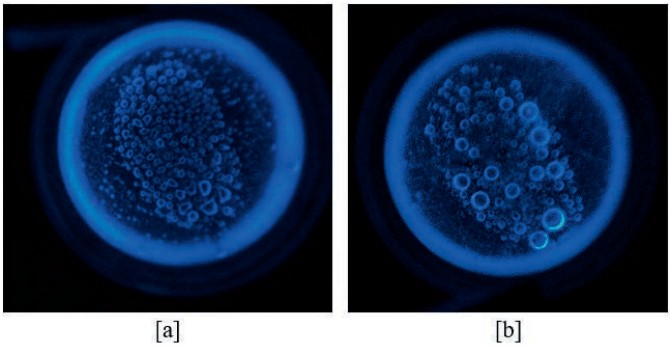


Fig. 4: Images created with ECL of fingerprints with vaseline adhesion.

The visualization of the eccrine fingerprints was not possible with our method. In subsequent processes, it is necessary to either pre-treat the eccrine impression fingerprint or to implement a different methodology in the prototype specifically for this evidence.

4 Discussion and future work

The observed peak in intensity right at the start of each experiment was caused by the output capacity of the used power supply creating a short burst in current. Utilizing this effect in the millisecond range requires an optimal timing between applying the current and image acquisition.
In summary, the visualization of sebaceous fingerprints and fingerprints with adhesions was possible with the presented method. The vaseline adhesion complicated the visibility

of anatomic features and also caused a bubble-like pattern which could be specific for this kind of adhesion and should be further investigated. The solubility of eccrine fingerprints prevented a successful visualization. It can be assumed that fresh fingerprints are better soluble in water than old ones. The ridges of the old trace are optically smaller than by the fresh one. As a result, it's easier to see the anatomic features. Our presented method therefore works better with traces that already dried up. Furthermore, fingerprints with other adhesions are to be investigated, preferably with criminally relevant background. Subsequent image processing may be one way to improve the results.
It should also be noted that there is a need to add hydrogen peroxide to the luminol solution, as it acts as a catalyst, even though it increases the formation of bubbles and should therefore be kept to a minimum.
The current approach is limited to smooth conductive surfaces. In future work transfering fingerprints from various surfaces to the ECL analysis cell should be tested. Future work should focus on the visualization of fingerprint adhesions. Forensic relevant information like gender and age can be determined from the ratio of different amino acids and fatty acids. Those adhesions can also serve as a hint for the usage of drugs or fire accelerators [AIA12, Du17, Gi16]. In summary, it is desirable to gain more information of a fingerprint than the anatomic features. ECL approves to be a good approach to reach this goal. Those information can be used to increase the succes rate of identification. The linkage between dactyloscopy and ECL can change forensic casework in terms of duration and quality, but needs further scientific analysis to develop its full potential.

5 Acknowledgement

This projekt was founded by the "Bundesministerium für Wirtschaft und Technologie (BMWi) - Fördermodul Koperationsprojekte (KF)" within the framework of the "Zentrales Innovationsprogramm Mittelstand" (ZIM).
We would like to thank all participants of the project partners of IVV Fraunhofer Dresden and the company Helling GmbH as well as the students Maria Izaber for their assistance.

References

[Ad11] Adcock, Jacqui L.; Barrow, Colin J.; Barnett, Neil W.; Conlan, Xavier A.; Hogan, Conor F.; Francis, Paul S.: Chemiluminescence and electrochemiluminescence detection of controlled drugs. Drug testing and analysis, 3(3):145–160, 2011.

[AIA12] Akinlabi, Esther T.; Isvarial, Milan; Akinlabi, Stephen A.: Design Of An Innovative Accelerant Detector. 2012.

[BH10] Beresford, Ann L.; Hillman, A. Robert: Electrochromic enhancement of latent fingerprints on stainless steel surfaces. Analytical chemistry, 82(2):483–486, 2010.

[Du17] Dunstan, R. H.; Sparkes, D. L.; Dascombe, B. J.; Stevens, C. J.; Murphy, G. R.; Macdonald, M. M.; Gottfries, J.; Gottfries, C-G; Roberts, T. K.: Sex differences in amino acids lost via sweating could lead to differential susceptibilities to disturbances in nitrogen balance and collagen turnover. Amino acids, 49(8):1337–1345, 2017.

[Ea11] Eagleson, Mary: Concise Encyclopedia Chemistry. De Gruyter, Hawthorne, 2011.

[FBK09] Forster, Robert J.; Bertoncello, Paolo; Keyes, Tia E.: Electrogenerated chemiluminescence. Annual review of analytical chemistry (Palo Alto, Calif.), 2:359–385, 2009.

[GA13] Gressner, Axel M.; Arndt, Torsten: Lexikon der Medizinischen Laboratoriumsdiagnostik. Springer, Berlin and Heidelberg, 2., überarbeitete und erweiterte auflage edition, 2013.

[Gi16] Girod, A.; Ramotowski, R.; Lambrechts, S.; Misrielal, P.; Aalders, M.; Weyermann, C.: Fingermark age determinations: Legal considerations, review of the literature and practical propositions. Forensic science international, 262:212–226, 2016.

[GRW12] Girod, Aline; Ramotowski, Robert; Weyermann, Céline: Composition of fingermark residue: a qualitative and quantitative review. Forensic science international, 223(1-3):10–24, 2012.

[He22] Heindl, Robert: System und Praxis der Daktyloskopie und der sonstigen technischen Methoden der Kriminalpolizei. De Gruyter, Berlin and Boston, 1922.

[HS07] Herrmann, Bernd; Saternus, Klaus-Steffen: Kriminalbiologie. Springer E-book Collection. Springer, Berlin, 2007.

[Ja15] Jasuja, O. P.; Singh, Kulvir; Kumar, Parveen; Singh, Gagandeep: Development of latent fingermarks by aqueous electrolytes on metallic surfaces: further studies. Canadian Society of Forensic Science Journal, 48(3):122–136, 2015.

[Ji13] Jiao, Tifeng; Huang, Qinqin; Xiao, Yong; Shen, Xihai; Zhou, Jingxin; Gao, Faming: Electrochemiluminescent Detection of Hydrogen Peroxide via Some Luminol Imide Derivatives with Different Substituent Groups. Journal of Chemistry, 2013:1–6, 2013.

[LZJ06] Li, Jianguo; Zhao, Fengjuan; Ju, Huangxian: Simultaneous determination of psychotropic drugs in human urine by capillary electrophoresis with electrochemiluminescence detection. Analytica chimica acta, 575(1):57–61, 2006.

[PS74] Periasamy, N.; Santhanam, K. S. V.: Studies on efficiencies of electrochemiluminescence of rubrene. Proceedings of the Indian Academy of Sciences - Section A, 80(4):194–206, 1974.

[Sp37] Specht, W.: Die Chemiluminescenz des Hämins, ein Hilfsmittel zur Auffindung und Erkennung forensisch wichtiger Blutspuren. Deutsche Zeitschrift für die Gesamte Gerichtliche Medizin, 28(1):225–234, 1937.

[Su16] Su, Bin: Recent progress on fingerprint visualization and analysis by imaging ridge residue components. Analytical and bioanalytical chemistry, 408(11):2781–2791, 2016.

[Va16] Valenti, Giovanni; Fiorani, Andrea; Li, Haidong; Sojic, Neso; Paolucci, Francesco: Essential Role of Electrode Materials in Electrochemiluminescence Applications. ChemElectroChem, 3(12):1990–1997, 2016.

[Xu12] Xu, Linru; Li, Yan; Wu, Suozhu; Liu, Xianghong; Su, Bin: Imaging latent fingerprints by electrochemiluminescence. Angewandte Chemie (International ed. in English), 51(32):8068–8072, 2012.

[YA04] Yager, Neil; Amin, Adnan: Fingerprint verification based on minutiae features: a review. Pattern Analysis & Applications, 7(1):94–113, 2004.

[YF11] Yamashita, Brian; French, Mike: Latent print development. The fingerprint sourcebook, pp. 7–67, 2011.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 255

ChildFace: Gender Aware Child Face Aging

Praveen Kumar Chandaliya[1], Aditya Sinha[2], Neeta Nain[3]

Abstract: Child face aging and rejuvenation has amassed considerable active research interest due to its immense impact on monitoring applications especially for finding lost/abducted children with childhood photos and hence protect children. Prior studies are primarily motivated to enhance the generation quality and aging of face images, rather than quantifying face recognition performance. To address this challenge we propose ChildFace model. Our model does child face aging and rejuvenation while using gender as condition. Our model uses Conditional Generative Adversarial Nets (cGANs), VGG19 based perceptual loss and LightCNN29 age classifier and produces impressive results. Intense quantitative study based on verification, identification and age estimation proves that our model is competent to existing state-of-art models and can make a significant contribution in identifying missing children.

Keywords: Child Face Aging, Generative Model, Face Recognition, Age estimation.

1 Introduction

Child trafficking is a grave problem world over. The National Centre for Missing and Exploited Children (NCMEC) [Naa], reports approximately $8,00,000$ children go missing in USA every year. The National Crime Records Bureau, India, reported the total human trafficking is $88,008, 6.9\%$ more than the previous year [Nab]. On average, the victim ratio is $1:6$ for boys to girls. Locating missing children over $[5-10]$ years time lapse is very complicated by the fact that child face changes dramatically as they age, making longitudinal face recognition [CE09] infinitely difficult. Our work can help law enforcement agencies and governments attempting to use cross-age face recognition system to find missing children and track other types of child exploitation. Our model namely, ChildFace consists of a generator G and 2 multi-scale discriminators denoted as D_1, D_2, LightCNN-29 [Wu18] based age classifier L_{age} and VGG19 [SZ15] based perceptual loss network L_{vgg}. Our key efforts are summarized as following:

1. We propose a new coarse-to-fine generator and multi-scale discriminator architecture for child age progression or regression considering age and identity as conditions.
2. Existing state-of-art face aging models does not address the challenge of child face aging and deep CNN based face recognition for quantitative evaluation of models. We present arguably the most intensive experimental evaluation with the help of state-of-art face matchers - FaceNet [SKP15], PFE [SJ19] and ArcFace [De19] on CLF Test dataset [CN19].

[1] Malaviya National Institute of Technology, Jaipur, India, 2016rcp9511@mnit.ac.in
[2] Malaviya National Institute of Technology, Jaipur, India, 2016ucp1447@mnit.ac.in
[3] Malaviya National Institute of Technology, Jaipur, India, nnain.cse@mnit.ac.in

3. Our proposed model ChildFace stands out to be second-best in rank-1 identification accuracy, verification and age estimation in comparison to CAAE, AIM and CPAVAE models.

2 Related Work

2.1 Age Progression and Regression

Face aging is the expectation of future looks and revival is the estimation of more youthful appearances, likewise alluded to as facial age relapse. It significantly affects a wide scope of utilization in different spaces. Generative adversarial network (GAN) [Go14] are continuously used to perform face aging and de-aging because of their undeniably creating conceivable and compelling faces with an antagonistic training approach. In this methodology, face aging and de-aging is ordered in two classes.

1. **Adversarial Auto Encoder:** In this architecture, we could feed age, gender (any other condition) along with random latent distribution as a one-hot conditional vector into the decoder to generate progressed and regressed age faces. Antipov *et al.* [ABD17] proposed an aging framework which also applies conditional GAN and used a local manifold adaptation (LMA) technique for identity approximation. Zhang *et al.* proposed CAAE [ZSQ17] for face aging and de-aging framework that learns a face manifold with adversarial training imposed on the encoder and generator, respectively, forcing to generate realistic faces. This model was not efficient in-term of face generation quality and recognition.
2. **Conditional Generative Adversarial Networks:** Imposing a one-hot condition vector along with image thereby feeding into encoder or generator of the model. PAG-GAN [Ya18a] designed pyramidal adversarial discriminator at multiple scales, which extracts high-level aging features. IPCGAN [WXTG18], is an identity preserved conditional GAN which functions as the face generator, and an age classifier forcing the face generation at the target age. Also, along similar lines, Sun *et al.* [Su20] proposed Label Distribution-guided GAN(IdGAN) to investigate age simulation over long-term and short-term aging sequence only on adult faces.

Majority of the aforementioned studies did not study face recognition performance whether it is identification, verification, age estimation and also experimented adult face aging and do not address child face aging.

3 Architecture of Proposed Model

The complete network architecture of the proposed model is illustrated in Figure 1. The model contains 2 trainable network components, namely Generator (G) and Multi-Scale Discriminator (D). Motivated from results of IPCGAN [WXTG18] we adopt their architecture of G. To learn gender aware age distribution we add gender condition - $T_G \in \mathbb{R}^{128\times128\times1}$ along with the age condition $T_A \in \mathbb{R}^{128\times128\times5}$. G receives child face image as input $x \in \mathbb{R}^{128\times128\times3}$ forming the final input as $(x, T_A \cap T_G) \in \mathbb{R}^{128\times128\times9}$. For Multi-Scale Patch Discriminator we use the architecture of IPCGAN[WXTG18] but instead of using single discriminator we use multi-scale patch discriminator i.e., two patch discriminators

Fig. 1: Detailed architecture of the proposed model.

(D_1 and D_2) discriminating at different image scales. For perceptual loss network we use pretrained VGG19 network. We extract the features from *relu*3_1 layer of the network. We then compare the high level features of input face and generated face in feature space. For age classification, we train LightCNN-29 model on CLF dataset. We predict the age of generated face, if the age is correct small penalty is given otherwise large penalty is given.

4 Loss Functions

Age transformation aims to generate a face ($\bar{x}$) that lies in target age group (T) while retaining the identity of (x). To achieve these goals, we train our model using adversarial loss, perceptual loss and age classification loss.

1. *Adversarial Loss:* As a standard process of GAN model, the generator G and discriminators (D_1 and D_2) are trained alternatively via an adversarial loss. D_1 and D_2 collectively try to distinguish real images at different scales (resolutions) from synthesized ones.

$$L_{adv}^{D} = \frac{1}{4}\sum_{i=1}^{2}\{\mathbb{E}_{x\sim p_x(x)}[(D_i(x|T_A \cap T_G)-1)^2] + \mathbb{E}_{y\sim p_y(y)}[(D_i(G(y|T_A \cap T_G)))^2]\}$$
$$L_{adv}^{G} = \frac{1}{4}\sum_{i=1}^{2}\{\mathbb{E}_{y\sim p_y(y)}[(D_i(G(y|T_A \cap T_G))-1)^2]\} \quad (1)$$

   As the optimization of Conditional Generative Adversarial Networks (cGAN)[MO14] suffers from instability and therefore the generated images have lot of artifacts. To improve the optimization of cGANs we adopt LSGANs [Ma16].
2. *Perceptual Loss*: The adversarial loss all alone can not guarantee that the synthesized faces retain the identity information. To address this we incorporate perceptual loss network using pre-trained VGG19 [SZ15] feature extractor, which increases the semantic similarity between input faces (x) and synthesized faces ($\bar{x}$) using Euclidean

distance in the deep feature space. Here, both x and $\bar{x}$ faces are sent to VGG19 which outputs their feature maps - $\Phi_{relu3\_1}(x)$ and $\Phi_{relu3\_1}(\bar{x})$ respectively from $relu3\_1$ layer. Perceptual Loss is defined as:

$$L_{vgg} = \frac{1}{2{\times}C \times W \times H} \sum_{c=1}^{C} \sum_{w=1}^{W} \sum_{h=1}^{H} \|\Phi(x)^{c,w,h}_{relu3\_1} - \Phi(\bar{x})^{c,w,h}_{relu3\_1}\|^2 \tag{2}$$

Here C, W, H denotes number of filters, width and height of the feature map respectively.

3. *Age Classification loss*: To achieve better age classification performance specifically on small dataset and to assure faster convergence, we train child dataset on light convolutional neural network (LightCNN29). We then use this network to determine age classification loss.

$$L_{age} = -\sum_{i=0}^{5} y_i \log(p_i) \tag{3}$$

Here y_i denotes the desired age group of generated face and p_i denotes the predicted age group of generated face by age classifier.

The final objective function is:

$$G_{loss} = \lambda_1 L^G_{madv} + \lambda_2 L_{vgg} + \lambda_3 L_{age} \tag{4}$$

$$D_{loss} = L^D_{madv} \tag{5}$$

Here λ_1, λ_2 and λ_3 are hyperparameters.

5 Experimental Setup

For training and evaluation of ChildFace model, we choose the Children Longitudinal Face (CLF) [CN19]. CLF contains $35,484$ child face images from $9,475$ longitudinally paired and $7,494$ single subjects, in the age group $[2:19]$ years annotated with age and gender. For the training of our model, we use $31,936$ images. For face verification and identification purpose we have created a separate test dataset namely, CLF Test dataset (identity-disjoint from training data) which includes 278 pair of youngest and oldest image of same subject with a time lapse of $[5-8]$ years. To train our model we divide the training data-set into 5 non-overlapping age groups, i.e., $[2{-}5]$, $[6{-}8]$, $[9{-}11]$, $[12{-}14]$, $[15{-}19]$.

5.1 Implementation Details

For data pre-processing, we used MTCNN [Zh16] to detect the five landmark points (two eyes, nose, and two mouth corners) that are used for proper alignment and to crop the images to a resolution of 128×128 pixels. Before passing images into model, they are normalized pixel-wise. For training ChildFace, all components are trained with a batch size of 32 using Adam optimizer with hyper-parameter $\alpha = 0.0001$ and $\beta = (0.5, 0.999)$. λ_1, λ_2 and λ_3 are set to 75.0, 1.0 and 1.0 respectively. The output of Generator G is restricted to $[-1, 1]$ using tanh activation function. The model is trained for $75K$ iterations. The model was trained from scratch with a learning rate of G, D_1, D_2 as 0.0001. We perform one optimization step for G and two optimization steps for D_1 and D_2.

6 Qualitative Evaluation

We do qualitative study of our model by comparing with 4 different face-aging models. Figure 2 shows that CAAE generates blurry and occluded face. IPCGAN produces good quality images compared to CAAE, aging performance is also good but there is change in contrast, and color composition of images. CPAVAE produces highly illuminated and smooth pictures for all faces which leads to identity loss as the texture of resultant face skin is enhanced drastically, as the model is based on deep feature consistency principle. In the case of AIM all faces look similar because age and identity component are disentangled. The reason is that AIM includes disentangled representation learning network. Compared to IPCGAN, CPAVAE and AIM, in our model only components responsible for aging are altered, whereas, noise factors such as skin color, quality, pose and background remain consistent which is compromised in former models for e.g., see subject 3 **(bottom left)** in Figure 2.

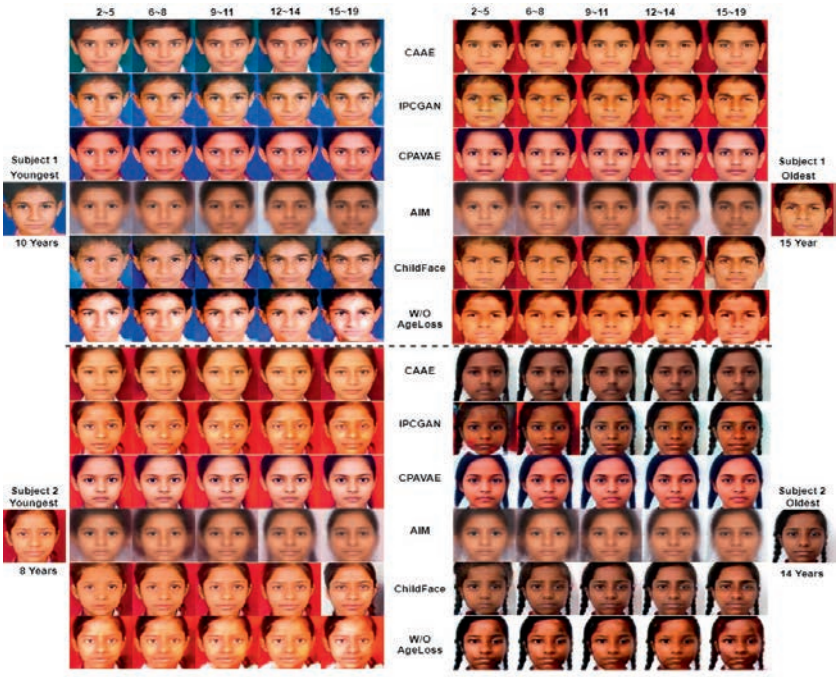
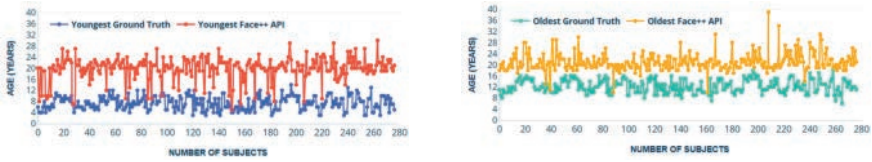


Fig. 2: Comparison with prior work on aging: CAAE, IPCGAN, CPAVAE, AIM and our ChildFace model, respectively. In the left-most and right-most column we show test faces and their ground truth ages.

7 Quantitative Evaluation

Previous face aging state-of-the-art frameworks [ABD17, ZSQ17, Ya18a, WXTG18, Zh19] work well for adult faces and long-span aging. These models use Face++ API for identity preservation and aging accuracy comparison. Our work is focused primarily for children, and Face++ tool does not work well on child datasets. To support this claim, we conduct an experiment on CLF Test dataset. We evaluated the age predicted by Face++ API. As we can see in Figure 3, the difference between the Youngest Ground Truth age and

Face++ API predicted age is in the range of [10-25] years and Oldest Ground Truth age and Face++ API predicted age is in the range of [5-25] years. This proves our claim that Face++ API does not give proper age approximation on child faces. To evaluate our model



(a) (b)

Fig. 3: Age estimation (years). A comparison between a) Youngest Ground Truth age Vs Face++ API and b) Oldest Ground Truth age Vs Face++ API predicted age on CLF Test dataset.

performance, we employ 3 open-source face matchers: FaceNet [SKP15] is trained on VGGFace2 dataset using the softmax loss. Probabilistic Face Embeddings (PFE) [SJ19] is trained using 64-CNN layers on MS-Celeb-1M dataset and an Additive Angular Margin Loss (ArcFace) [De19] is trained on MS-ArcFace dataset. All aforementioned matchers extract a 512 dimensional embedding feature.

7.1 Face Verification

For face verification we have used two genuine image pair protocols. 1) The youngest ground truth picture is compared to the oldest ground truth GT picture of the same subject-Real Young vs Real Old Pair(RY-RO). This test is to compare the variance of the actual with the synthesized images. 2) The oldest GT picture is compared with the synthesized aged picture of the same age group generated by the proposed model - Real Old vs Artificial Old Pair(RO-AO). The cross-subject imposter (negative) pairs are created by randomly pairing the progressed or regressed images and real images of individual subjects. In total, there are 278 genuine matches and $7,146$ imposter matches.

1. IPCGAN and ChildFace models only progress/regress the age retaining the identity. Due to this reason the verification accuracy is high (see Table 2).
2. CPAVAE, AIM, CAAE face aging models generate age progress and regress face by latent vector approximation, due to which identity is also altered along with age. Thus, the face verification accuracy is low compared to IPCGAN and ChildFace.
3. RY-RO verification accuracy is low compared to RO-AO. In the first case youngest and oldest faces are compared. On the other hand, in RO-AO both the real and generated faces of old age group are compared. Therefore, the Face verification in latter protocol is high.

7.2 Face Identification

In this paper, the performance is reported in terms of rank-1 closed-set identification accuracy (recovered child is in the gallery) under the youngest (gallery) to oldest (probe) protocol. For all our experiments, the gallery set has 278 faces from different subjects with generated (5 images) of each subject, while 278 (oldest) images in probe set. In Table 1, we report the rank-1 accuracy of our model and state-of-the-art models on synthesized age progressed images. We find that our model achieves second-best search accuracy compared to IPCGAN [WXTG18].

Tab. 1: Face verification and identification performance (%) of different face aging models using CNN-based face recognition models on CLF Test dataset. Face Match rates at FMR=0.1%. Bold show best score, and bold italics show second-best score.

Model	**FaceNet** [SKP15]			**PFE** [SJ19]			**ArcFace** [De19]		
	GMR@FMR (RY-RO)	**GMR@FMR** (RO-AO)	**Rank-1%**	**GMR@FMR** (RY-RO)	**GMR@FMR** (RO-AO)	**Rank-1%**	**GMR@FAR** (RY-RO)	**GMR@FMR** (RO-AO)	**Rank-1%**
CAAE [ZSQ17]	16.55	3.25	6.45	63.67	3.50	6.88	86.33	6.11	8.33
IPCGAN [WXTG18]	**28.78**	**99.28**	**74.82**	**78.78**	**99.98**	**92.45**	**90.65**	**99.28**	**97.12**
CPAVAE [CN19]	15.47	5.76	21.22	62.95	4.67	20.86	88.49	2.07	21.22
AIM [Zh19]	***25.51***	5.04	28.42	79.50	14.03	56.12	90.65	5.04	54.68
ChildFace	24.10	***97.12***	***71.58***	***79.85***	***99.64***	***90.29***	***94.97***	***94.25***	***93.16***

7.3 Age Estimation

Identically, objective age estimation is conducted to measure the aging and de-aging accuracy. We employ state-of-the-art publicly available CNN based age estimation model, namely, Soft Stagewise Regression Network (SSR-Net) [Ya18b]. Table 2 shows the age estimation results on progressed or regressed images by state-of-the-art and proposed model. ChildFace age estimation in age range $[2-8]$ is far better compared to other models.

Tab. 2: Age Progression and Regression: Objective age estimation (years) evaluated by SSR-Net [Ya18b]. Due to the limited space, we only address the mean and standard deviation of age estimation computed over CLF Test dataset.

Age group	Age Estimation (year) of CLF Test dataset				
	2-5	**6-8**	**9-11**	**12-14**	**15-19**
CAAE [ZSQ17]	8.76 ± 5.01	11.72 ± 5.89	14.62 ± 6.23	$\mathbf{16.83 \pm 6.07}$	$\mathbf{19.33 \pm 5.42}$
IPCGAN [WXTG18]	6.40 ± 5.30	10.92 ± 6.85	17.16 ± 8.11	20.07 ± 8.47	19.75 ± 8.08
CPAVAE [CN19]	8.84 ± 5.54	11.61 ± 6.42	$\mathbf{14.21 \pm 6.47}$	17.83 ± 6.19	19.55 ± 5.62
ChildFace(WithAgeLoss)	$\mathbf{5.49 \pm 5.62}$	$\mathbf{11.28 \pm 6.44}$	18.29 ± 8.24	22.00 ± 7.26	25.81 ± 5.82
Without AgeLoss	18.26 ± 7.41	16.21 ± 7.94	14.71 ± 7.95	11.15 ± 7.65	13.52 ± 8.47

8 Ablation

For analyzing the effect of age classification loss. Figure 2 present the visual comparisons between ChildFace and Without Age loss function which penalizes aging. We observed that generated faces by ChildFace are more realistic. Furthermore, Table 2 shows that ChildFace outperforms without Age loss in age estimation. While with Age loss function, more aging effect and clear texture information is produced in all age groups.

9 Conclusions

We propose a conditional GAN-based model to solve age progression and regression on child faces. Compared to previous approaches this model specifically does short-span aging by employing multi-scale patch discriminators (better critic), LightCNN-29 Networks and pretrained VGG19 Network. We report intensive performance evaluation through comparison with 3 state-of-art matchers. Experiment results show that the proposed model can improve the ability to locate and identify young children who are possible victims of abduction or child trafficking. In future, we intend to extend our research on unconstrained child face aging and recognition with improvement in accuracy.

10 Acknowledgments

This research is based upon work supported by the Ministry of Electronics and Information Technology (MeitY), Government of India, under Grant (No.4 (13)/2019-ITEA). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the TITAN V GPU used for this research.

References

[ABD17] Antipov, Grigory; Baccouche, Moez; Dugelay, Jean-Luc: Boosting cross-age face verification via generative age normalization. In: IEEE. BTAS, 2017.

[CE09] Cunha E, Baccino E, Martrille L Ramsthaler F Prieto J Schuliar Y Lynnerup N Cattaneo C.: The problem of aging human remains and living individuals: A review. Forensic Sci Int, 2009.

[CN19] Chandaliya, Praveen Kumar; Nain, Neeta: Conditional Perceptual Adversarial Variational Autoencoder for Age Progression and Regression on Child Face. In: ICB. pp. 1–8, 2019.

[De19] Deng, Jiankang; Guo, Jia; Niannan, Xue; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR. 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: NIPS. 2014.

[Ma16] Mao, Xudong; Li, Qing; Xie, Haoran; Lau, Raymond Y. K.; Wang, Zhen: Multi-class Generative Adversarial Networks with the L2 Loss Function. CoRR, 2016.

[MO14] Mirza, Mehdi; Osindero, Simon: Conditional Generative Adversarial Nets. CoRR, 2014.

[Naa] National centre for Missing and Exploited Children. http://www.missingkids.com.

[Nab] National Crime Record Bureau:. http://ncrb.gov.in.

[SJ19] Shi, Y.; Jain, A.: Probabilistic Face Embeddings. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6901–6910, 2019.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: CVPR. 2015.

[Su20] Sun, Y.; Tang, J.; Shu, X.; Sun, Z.; Tistarelli, M.: Facial Age Synthesis With Label Distribution-Guided Generative Adversarial Network. IEEE Transactions on Information Forensics and Security, 15:2679–2691, 2020.

[SZ15] Simonyan, Karen; Zisserman, Andrew: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR. 2015.

[Wu18] Wu, Xiang; He, Ran; Sun, Zhenan; Tan, Tieniu: A light CNN for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security, pp. 2884–2896, 2018.

[WXTG18] Wang, Z.; X. Tang, W. Luo; Gao, S.: Face Aging with Identity-Preserved Conditional Generative Adversarial Networks. In: CVPR. 2018.

[Ya18a] Yang, Hongyu; Huang, Di; Wang, Yunhong; Jain, Anil K.: Learning Face Age Progression: A Pyramid Architecture of GANs. In: CVPR. pp. 31–39, 2018.

[Ya18b] Yang, Tsun-Yi; Huang, Yi-Hsuan; Lin, Yen-Yu; Hsiu, Pi-Cheng; Chuang, Yung-Yu: SSR-Net: A Compact Soft Stagewise Regression Network for Age Estimation. In: IJCAI-18. pp. 1078–1084, 2018.

[Zh16] Zhang, Kaipeng; Zhang, Zhanpeng; Li, Zhifeng; Qiao, Yu: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. pp. 1499–1503, 2016.

[Zh19] Zhao, Jian; Cheng, Yu; Cheng, Yi; Yang, Yang; Lan, Haochong; Zhao, Fang; Xiong, Lin; Xu, Yan; Li, Jianshu; Pranata, Sugiri et al.: Look Across Elapse: Disentangled Representation Learning and Photorealistic Cross-Age Face Synthesis for Age-Invariant Face Recognition. AAAI, 2019.

[ZSQ17] Zhang, Zhifei; Song, Yang; Qi, Hairong: Age Progression/Regression by Conditional Adversarial Autoencoder. In: CVPR. pp. 4352–4360, 2017.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 265

Effects of sample stretching in face recognition

Mathias Fredrik Hedberg[1]

Abstract: Face stretching is something that can occur intentionally and unintentionally when preparing a face sample for enrollment in a face recognition system. In this paper we assess what affects both horizontal and vertical stretching have on a face recognition algorithms. Basic closed-set identification tests revealed that holistic face recognition algorithms performed poorly compared to feature based recognition algorithms when classifying non-stretched samples against templates based on stretched samples.

Keywords: Face recognition, presentation effects.

1 Introduction

Face recognition technology has proven itself to be an accurate and relatively non-intrusive method for biometric recognition. As face recognition is implemented in more important aspects of our society, it is critical that theses systems perform as expected. In this paper we assess the effects of presenting stretched face images during the enrollment stage, and how this affects future classifications.

Stretching of samples may occur both unintentionally (such as from printer or scanner issues) and in some cases intentionally (possibly to avoid recognition or even for beautification). The U.S Department of state estimates that around 12% of online via applications are stretched and recognizes that stretched images can severely impact the accuracy of face recognition [MM18]. The consequences of this could result in a subject applying for multiple visas under different names with the same face, breaking the integrity of such an application process.

For this paper we will narrow the scope to electronic identity documents where the enrollment sample has been stretched either vertically or horizontally. To achieve an adequate understanding of the key elements of this topic, a current state-of-the art literature study is presented, with focus directed on the impacts of stretched face images in face-recognition systems. Finally, a practical experiment is conducted to measure said impacts on modern face recognition systems.

[1] Norwegian University of Science and Technology, Gjøvik, mathifhe@stud.ntnu.no

2 State-of-the-art

2.1 Image stretching

In the realm of face recognition, there are many scenarios where some form of image stretching can occur, resulting in a distorted image being presented to the system. One major issue is the result of using wide angle lenses to increase the field of view. This creates what is known as barrel distortion, which causes curvature around the periphery of an image. Captured faces located in this area are more challenging to recognize, however successful methods have been developed to mitigate such distortions where the capture environment is known [JNK08].

Next we have the concept of vertical contraction and extension. This type of stretching is the result of one axis being fixed, while the other is either extended or contracted, creating a distortion of the aspect ratio of the image. Take note that vertical contraction yields the same result as horizontal extension, and vertical extension the same as horizontal contraction (when not taking image size into account).

A 2014 study [Su14] was conducted to measure the recognition impact of face alterations on face recognition systems, in addition to detection of such alterations. The study made use of three different face recognition solutions which at the time were state-of-the-art. These were Neurotchnology VeriLookSDK (VL), Luxland FaceSDK (LU) and a SIFT-based matching algorithm (SI). The study found that some systems such as LU and SI were highly sensitive to alterations of the aspect ratio, while VL saw no recognizable performance impacts. The author of the study believed that this was due to VL being based on feature-based recognition, and that feature-based recognition is quite insensitive to global geometric changes.

2.1.1 Detecting stretched images

Detecting if a face sample has been stretched can be done in two ways, analyzing the sample directly, or comparing the presented sample to a controlled non-stretched sample. Analyzing the sample directly with no other references is increasingly difficult as stretch magnitude decreases [MM18]. The US Department of State has done some research in this field [MM18] making use of Convolutional Neural Network (CNN) to detect stretching based on training the network on artificially stretched visas application images. Based on a Layer-wise Relevancy Propagation (LPR) analysis of the network, they found that LPR maximums appeared within the eyes. This can boil down to the fact that most pupils are usually circle shaped, and an oval shaped pupil could indicate that stretching has taken place.

Comparing a stretched sample to a non-stretched sample of a subject face is not always possible, however it does make stretched face image detection easier. The 2014 study [Su14] was also successful in detecting stretching in 90% of face images where the stretch

intensity was only 10%. In this case the detection required that the subject was present for comparison against the enrollment sample.

2.2 Image normalization

Image normalization is a process that takes place in most face recognition pipelines after face detection to eliminate unwanted environmental variables. There are many approaches to this, many algorithms focus on the eyes exclusively when doing geometric alignment, scaling and rotation of the image so that the placement of eyes is the same in all images [CK10] [Li06]. Normalizing a stretched face however, may result in features such as the nose, mouth and ears having different locations compared to the same non-stretched face. Face recognition systems that deal with stretched samples must be able to tackle such variations if they are to perform well with stretched samples.

2.3 Feature extraction

Feature extraction is a critical element in the face recognition pipeline. This involves representing the face by extracting the most relevant details so that a face template can be created for future classification tasks. There are a range of approaches to this, however in general this stage can be divided into three categories; feature-based-, holistic- and hybrid-models [Zh03].

Holistic models process and represent the face image as a whole, often focusing on variations in texture to assist in analyzing the face [ZF14]. One of the more popular approaches to representing the face is with eigenfaces based on Principal Component Analysis (PCA), which calculates the eigenvectors and eigenvalues from the co-variance matrix of the face images, extracting only the principal components resulting in a huge dimension reduction of the images [TP91]. A different well known algorithm is Local Binary Patterns Histograms (LBPH) which is a simple texture operator for holistic face representation.

Feature-based models (often referred to part based) often decompose the face image into blocks as opposed to holistic methods where the image is processed as a whole [Ma11]. Local features such as the mouth, eyes and nose are extracted and local statistics (geometric and/or appearance) are fed into a structural classifier [Zh03]. Feature-based models are less sensitive to variations in illumination, viewpoint and also any inaccuracies from the face-localization stage. One of the most recognized feature-based models is the CNN approach. CNN provides partial invariance to translation, rotation, scale, and deformation [La97]. The point of deformation is important for us as this indicates that CNN may cope well with sample stretching. Histogram of Oriented Gradients (HOG) is also a very well known feature-based model that functions similarly to CNN [DT05].

Based on this information, it can be assumed that processing stretched and non-stretched face samples through holistic and feature-based recognition systems will result in better classifications on feature-based systems, mainly due to the heavy geometric reliance found in holistic algorithms.

3 Practical experiment

A practical experiment was proposed in the form of a scenario evaluation to help get a better understanding of the presentation effects of face stretching. To achieve this, a database of face samples was retrieved and a selection of enrollment samples were stretched in a controlled manner before being enrolled into various face recognition systems. The resulting templates were then used by these systems to classify multiple non-stretched samples of the same subjects.

This experiment design was chosen to best reflect the scenario of identity document creation. The stretched samples represent the images enrolled for use in the identity documents, while the non-stretched samples represent the real face representations that an operator would be presented with when verifying document ownership such as at an airport.

Testing was done on a consumer desktop workstation running Fedora 31 comprised of an Intel i5-7600K @3.80GHz x4 and an NVIDIA GTX 1060 3GB. dlib was compiled with CUDA support to make use of the GPU when available. Running the entire experiment consumed over 2 hours of processing time, however this could be further reduced by also compiling OpenCV with CUDA support, along with other code optimizations.

3.1 Stretching of face samples

This experiment made use of the University of Essex face database (faces94)[3] containing 395 subjects photographed 20 times. These samples were taken in a controlled environment over a very short period of time, including only front-facing face images with minimal pose variations and good lighting. Most of the samples may in fact adhere to the ISO/IEC 19794-5 standard for identity documents. This uniformity makes the database relatively easy for face recognition systems to work on, and the changes from any external presentation effects that are applied should be easier to detect. One enrollment sample was extracted for each individual while the rest of the samples were reserved for later identification testing.

(a) -20% stretch (b) 0% stretch (c) +20% stretch

Fig. 1: Result of ffmpeg stretching, while retaining the original 180x200 format

Stretching of face samples was achieved using the popular ffmpeg program using the 'scale' and 'pad' command line arguments. The aim of this stretching was to modify the

[3] `http://cswww.essex.ac.uk/mv/allfaces/index.html`

aspect ratio of the images without up-scaling pixels or modifying the image resolution. Letter-boxing (known as 'pad' in ffmpeg) is required to preserve the the pixel resolution of the container image. Figure 1 shows some example results from this stretching. A negative stretch intensity results in vertical contraction while a positive value results in horizontal contraction.

For this experiment, a total of 9 stretching values were used, from -40% to +40% at increments of 10%. This stretching was applied to each of the enrollment samples resulting in a total of 3555 enrollment samples (395 subjects at 9 stretch values). Stretching in excess of +-40% seems unrealistic in the presented scenario and would likely be detected by the operator without much effort, therefore no further stretching intensities were used. The stretch interval of 10% was chosen to retain a manageable processing time, but should ideally be lower.

3.2 Recognition software

Recognition was done using the standalone OpenCV library, and also using a newer face recognition library[4] based on the dlib toolkit. Software was written based on these libraries to perform closed-set identification using the following forms of recognition:

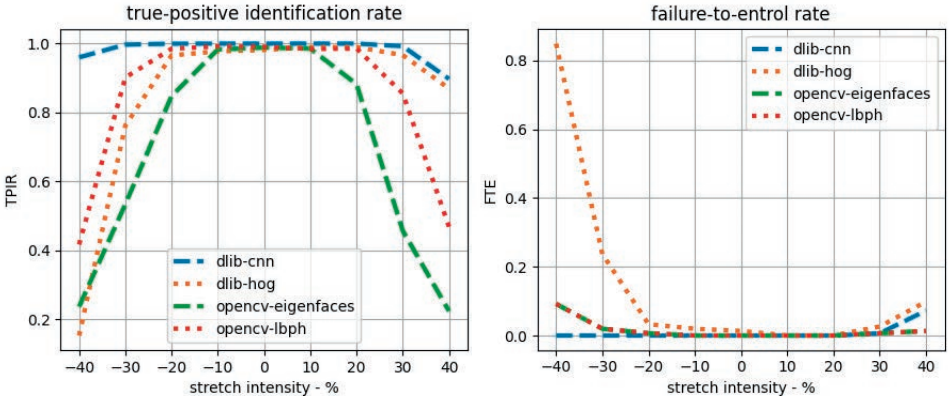
1. Eigenface based face recognition using OpenCV (holistic)
2. LBPH based face recognition using OpenCV (holistic)
3. CNN based face recognition using dlib (feature-based)
4. HOG based face recognition using dlib (feature-based)

As the software was to perform closed-set identification, there was no threshold set to identify impostors, the software was also therefore designed to make a decision even when confidence was low, as long as a face was detected. The OpenCV-based systems made use of Haar-cascade for face detection while the dlib-based systems used either HOG or CNN for face detection. Haar-cascading for face detection was currently unavailable in the dlib based recognition library used. The code used in this experiment can be found online[5].

3.3 Results

Figure 2a shows the true-positive identification rate (TPIR). The feature-based recognition systems (CNN and HOG) performed well in all stretching intensities between -20% and +30%. The holistic based methods (LBPH and eigenface) performed well between -10% and +10% stretch intensity, with LBPH extending this to +-20%. HOG had a sharp decline in TPIR when vertical contraction was more than 20% (-20% stretch intensity) compared to CNN which maintained results over 0.9 throughout all stretch intensities.

[4] `https://github.com/ageitgey/face_recognition`
[5] `https://github.com/metrafonic/FaceStretchCode`



(a) The true-positive identification rate

(b) The failure-to-entrol rate

Fig. 2: TPIR and failure to enroll (FTE) from 40% vertical contraction to 40% horizontal contraction

The FTE rate in this testing was all due to failures in face-detection in the enrollment stage. Observing the FTE rate in shown in figure 2b reveals that Haar-cascade based face detection used in the OpenCV based systems resulted in an FTE rate of less than 0.1 for all stretch intensities, with the CNN based recognition performing similarly. FTE was highest in the HOG based face detection, especially when enrolling samples that were vertically contracted.

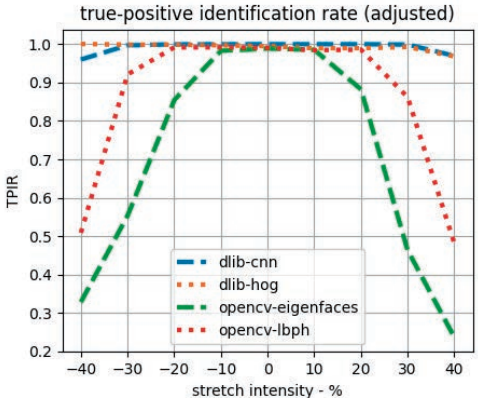


Fig. 3: The TPIR adjusted for FTE

The TPIR adjusted for when enrollment was successful is shown in figure 3. This graph shows that the feature-based recognition methods had a TPIR of more than 0.9 for all stretching intensities when subject enrollment was successful. The holistic based recognition models still showed poor performance for stretch values in excess of +- 30%. Eigenface recognition had the worst performance, seeing a sharp decline when working with samples stretched in excess of +- 20%.

4 Discussion

The results seen in the experiment show that holistic based recognition methods perform poorly when presented with samples that have been vertically or horizontally stretched. The feature based methods performed well throughout testing as long as face detection was achieved. This correlates with the findings mentioned earlier in [Su14]. It is likely that the feature based methods could perform even better in this experiment if Haar-cascading was implemented to assist in the face detection, as this gave good results in the OpenCV implementations used.

The poor performance in the eigenface recognition likely stems from the weaknesses of using face representations for classifications instead of performing classifications based on extracted features. LBPH takes a slightly different approach to face representation than eigenfaces giving it some headway, however it still suffers when stretching is over +- 30%. A fisherface based recognition was originally also a part of the systems tested, however it performed identically to the eigenface recognition as there was only a single training sample per subject, so the advantages of increasing between-class versus within-class scatter could not be harnessed. Due to this, the fisherface based algorithm was not used, however if more samples were to be used for enrollment, then fisherface based recognition may have an advantage compared to eigenface recognition.

Recognition performance was quite substantial when using feature-based models, though as this experiment was based on closed-set identification, there are some other untested factors that could severely impact the performance of the feature based algorithms. For this experiment, all the enrolled subject templates were of the same stretch value. Open-set identification that includes templates of other non-stretched samples may have impacted the classification in a negative manner.

The sharp increase in the FTE rate for HOG when handling samples stretched in excess of -40% had some unintended side-effects, and could be used to aid in sample rejection when presented with stretched samples.

Not mitigating presentation effect such as sample stretching, is essentially adding new variables to face recognition that could be avoided. Some variables like age are uncontrollable, but others such as tilt, lighting and pose variations are kept as uniform as possible. Sample stretching should not be treated any differently.

In the context of identity documents, these results highlight the importance of test identification and verification as a function of the enrollment process. If document issuers are avoiding the verification function, they are then essentially exposing themselves to a range of presentation attacks. As shown in [Su14], detection of 90% of stretched samples at over +-10% stretching is possible if a controlled reference sample is available. Implementing such checks at the enrollment location could help ensure the integrity of the enrollment samples.

5 Conclusion

Stretched face samples can severely impact the performance of some face recognition algorithms. Holistic recognition such as those making use of eigenfaces or LBPH are especially vulnerable when a face sample is vertically or horizontally stretched in excess of 20%. The feature based algorithms were relatively unaffected by sample stretching as long as face detection was achieved. Although feature based recognition performed well, different scenarios than those done here could yield other results. Further testing should be done with open-set identification scenarios to explore some of these potential weaknesses.

Countries issuing electronic Machine Readable Travel Document (eMRTD)s should verify that the enrollment sample has not been modified. This could be achieved by using some form of stretch detection on the presented sample by comparing it to a controlled sample, or by doing the whole image acquisition themselves instead of allowing citizens to capture them.

References

[CK10] Chaudhari, S. T.; Kale, A.: Face Normalization: Enhancing Face Recognition. In: 2010 3rd International Conference on Emerging Trends in Engineering and Technology. pp. 520–525, 2010.

[DT05] Dalal, Navneet; Triggs, Bill: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). volume 1. IEEE, pp. 886–893, 2005.

[JNK08] Jae Suhr, Kyu; Noh, Dong Hyun; Kim, Jaihie: Toward Face Recognition by Using a Fish-eye Camera. Korean Institute of Electrical Engineers, pp. 963–964, 2008.

[La97] Lawrence, S.; Giles, C. L.; Ah Chung Tsoi; Back, A. D.: Face recognition: a convolutional neural-network approach. IEEE Transactions on Neural Networks, 8(1):98–113, 1997.

[Li06] Li, G.; Cai, X.; Li, X.; Liu, Y.: An Efficient Face Normalization Algorithm Based on Eyes Detection. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3843–3848, 2006.

[Ma11] Marcel, Dr Sebastien: , An Introduction to Automatic Face Recognitionusing Statistical Models, 12 2011.

[MM18] McGarry, Delia; Melsom, Stephen: , Image Manipulation Detection & Effects of Perspective Distortion on Face Identification, November 2018.

[Su14] Sun, Yunlian: Advanced Techniques for Face Recognition under Challenging Environments. PhD thesis, University of Bologna, 2014.

[TP91] Turk, Matthew; Pentland, Alex: Eigenfaces for Recognition. J. Cognitive Neuroscience, 3(1):71?86, January 1991.

[ZF14] Zafaruddin, G. M.; Fadewar, H. S.: Face recognition: A holistic approach review. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). pp. 175–178, 2014.

[Zh03] Zhao, W.; Chellappa, R.; Phillips, P. J.; Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Comput. Surv., 35(4):399–458, December 2003.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 273

On the assessment of face image quality based on handcrafted features

Olaf Henniger[1] Biying Fu[1] Cong Chen[2]

Abstract: This paper studies the assessment of the quality of face images, predicting the utility of face images for automated recognition. The utility of frontal face images from a publicly available dataset was assessed by comparing them with each other using commercial off-the-shelf face recognition systems. Multiple face image features delineating face symmetry and characteristics of the capture process were analysed to find features predictive of utility. The selected features were used to build system-specific and generic random forest classifiers.

Keywords: Biometrics, face recognition, face image quality.

1 Introduction

Not all biometric samples are equally well suited for the automated recognition of individuals. The utility of a biometric sample, i.e. its usefulness for telling mated and non-mated samples apart, can be expressed by a quality score [ISO16]. The quality score can be used, e.g., for deciding whether the re-acquisition of data is deemed necessary, for weighting partial results in multi-biometric systems, or for selecting the best (in some sense) from a series of captured biometric samples. The utility of a biometric sample depends on its faithfulness to its source (i.e. fidelity) and the distinctiveness of the biometric features (i.e. character) [ISO16]. The utility of a biometric sample can be quantified after comparing it with mated and non-mated samples from a dataset. Hence, utility depends on the underlying dataset and on the feature extraction and comparison algorithms used.

Fields holding biometric sample quality scores were introduced into several standardized biometric data interchange formats [ISO19]. In these data formats, if a quality score is reported, valid values are integers between 0 and 100. According to [ISO18], quality scores from 0 to 25 should indicate unacceptable quality, from 26 to 50 marginal quality, from 51 to 75 adequate quality, and from 76 to 100 excellent quality. The calibration of the boundaries between the levels of quality is a considerable challenge.

Related work on predicting the utility of biometric samples concentrated on finger images [TWW04, T+16, ISO17] and iris images [TGS11, ISO15]. There is no standard yet for how to assess face image quality. To better understand face image quality assessment, algorithms can currently be submitted to NIST for evaluation [G+20]. A Technical Report

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany,
{ olaf.henniger | biying.fu }@igd.fraunhofer.de
[2] Technical University of Darmstadt, Department of Computer Science, Darmstadt, Germany

[ISO10] outlines a set of face image features that could be useful in calculating quality scores, but does not specify how to aggregate the individual feature values. It is currently under revision and amenable to contributions. A Technical Report on portrait quality [ICA18] includes requirements and recommendations on how to capture suitable reference face images, but does not specify how to assess the quality of face images captured in an arbitrary environment. Tools for automatically checking compliance to the ICAO requirements compute a number of individual scores, which however are not aggregated into an overall quality score [F+12].

For assessing face image quality, proprietary algorithms trained for particular face recognition algorithms are in use. This paper investigates how to predict the utility of face images across multiple state-of-the-art face recognition algorithms. We followed a supervised machine-learning approach similar to the one applied in finger image quality assessment [T+16, ISO17]. The goal was to learn a mapping from a face image feature vector to a scalar quality score. In contrast to [HO+19], which is about face image quality assessment using automatically learned features, we took "handcrafted" features into consideration, which were drawn from [ISO10]. The strength of handcrafted features is their explainability, which helps avoid using features with a potential demographic bias.

The rest of this paper is organized as follows: Section 2 describes the data available for this study. Section 3 deals with the a-posteriori assessment of the quality of a biometric sample by comparing it with other samples. Section 4 considers the a-priori quality assessment, i.e. predicting the utility of a sample without comparing it first with other samples. Section 5 assesses the accuracy of the utility-prediction model using a testing dataset. Finally, Section 6 summarizes the results and gives an outlook on future research work.

2 Underlying data

2.1 Face image dataset

The publicly available NIST Special Database 32 [C+09] was used in the analysis. It consists of 712 face images of 380 different test subjects. 686 images are frontal or nearly frontal face images. 26 images are full-profile or nearly full-profile images. For 145 test subjects, the dataset contains more than one different images, and for 69 test subjects even more than two different images. For one test subject, the dataset contains three identical images, only one of which was used in this study to avoid bias. For the other test subjects, there is only one image. The sizes of the uncropped frontal images range from 240×240 pixels to 1824×1170 pixels. The majority of images is of size 480×600 pixels.

2.2 Similarity scores

Two commercial off-the-shelf face identification systems were used to calculate similarity scores. In the following, they are referred to as System 1 and System 2. The systems are

treated as "black boxes" as we do not target at a comparative technology evaluation. The systems underwent the deep-learning revolution and successfully participated in NIST's recent Face Recognition Vendor Tests [GNH18a, GNH18b]. Both systems were configured so that each search returned 100 candidates, for whom the similarity scores were logged.

For both systems, all and only frontal face images were attempted to enrol into the reference database. Only frontal face images were enrolled because many face recognition systems do store frontal face images as reference images (e.g. in ePassports, forensic databases, entry/exit systems). System 1 encountered three failures to enrol. System 2 encountered no failures to enrol.

After enrolment, for each system, the reference database was searched against all (frontal and profile) face images. For utility assessment, comparisons of images with themselves were not taken into consideration. System 1 encountered three failures to extract, for the same images for which it encountered failures to enrol. System 2 encountered no failures to extract. For all frontal probe images without failure to extract, both systems returned all mated reference identifiers in the candidate lists. For profile probe images, neither system returned all mated reference images in the candidate lists. For this lack of mated similarity scores for profile images, we limited the study exclusively to the (nearly) frontal face images. Taking into account only frontal images, both systems returned all mated reference identifiers at the head of their candidate lists, i.e. all mated similarity scores were greater than any non-mated similarity score. Hence, despite their diversity, all frontal images in the dataset, except for the ones with a failure to extract, could be regarded as excellent quality with respect to state-of-the-art face recognition systems.

2.3 Training subset and testing subset

We partitioned the face image dataset randomly into nearly equally large disjoint training and testing subsets, leaving the subsets of face images for the same test subject undivided. The training dataset consisted of 345 face images of 190 test subjects. The testing dataset consisted of 339 face images of 190 test subjects.

3 A-posteriori assessment of utility

The utility of a biometric sample can be predicted in several ways, e.g.:

- For NIST's fingerprint image quality (NFIQ) assessment algorithm, version 1, the utility of a biometric sample was defined as the normalized difference between the mean of the sample's mated non-self similarity scores and the mean of the sample's non-mated similarity scores. Predicting a real-valued scalar is a regression problem. However, as regression methods failed to give adequate predictions, for NFIQ 1.0, the machine-learning problem was restated in terms of classification into utility classes (excellent, very good, good, fair, or poor utility) [TWW04].

- For the new and improved version NFIQ 2.0, a random decision forest was trained for binary classification into two utility classes (high or low utility). The trained random decision forest outputs class membership along with its probability. The quality score is the probability that an image belongs to the high-utility class multiplied by 100 and rounded to the nearest integer [T+16, ISO17].

The normalized difference between the mean of a sample's mated non-self similarity scores and the mean of its non-mated similarity scores can be calculated only if a sufficient number of randomly distributed mated and non-mated similarity scores were available. However, because similarity scores were available only for the most similar candidates, we chose another measure of separation of mated and non-mated similarity score distributions. For each frontal face image with more than one mated non-self similarity score, we computed a utility score as the normalized difference between the arithmetic mean of mated non-self similarity scores and the maximum non-mated similarity score. Images with only one mated non-self similarity score were not considered because the arithmetic mean of mated non-self similarity scores would be the same for both compared images, independent of their quality.

Using the training dataset, we built regression models between the utility score and multiple features specified in Section 4.1. However, these models failed to give adequate predictions of utility in the testing dataset. Therefore, like NFIQ 2.0, we tried binary classification into two utility classes. We selected face images of high and low quality as follows:

1. Class 1: All images whose minimum mated score was greater than the threshold value that corresponds to FNMR = 60% were labelled as high quality.
2. Class 0: All images whose maximum mated score was less than the threshold value that corresponds to FNMR = 30% were labelled as low quality.

The boundaries are arbitrary. They were chosen so that in the given training dataset about 40 images were of Class 1 and about 40 images were of Class 0 for either system. The remaining images neither belong to Class 1 nor to Class 0.

For each face recognition system, a specialized quality prediction model can be constructed. However, it would be useful to build a generic face image quality assessment model independent of particular face recognition systems. For this purpose, we formed unions and intersections of the Classes 1 and 0 of System 1 and System 2, respectively:

- The union of the Class 1 training sets consisted of 47 images that were of high quality for either System 1 or System 2. The union of the Class 0 training sets consisted of 49 images that were of low quality for either System 1 or System 2.
- The intersection of the Class 1 training sets consisted of 25 images that were of high quality for both System 1 and System 2. The intersection of the Class 0 training sets consisted of 36 images that were of low quality for both System 1 and System 2.

4 A-priori assessment of sample quality

4.1 Selection of face image features

Several face image features that could be suitable for predicting utility were proposed in [ISO10]. We coded the feature extraction in Python and extracted a feature vector consisting of the following elements from each face image:

- left-right (lighting and pose) symmetry [GLZ07, ISO10] calculated as sum of differences of normalized pixel luminance values of the left and right halves of the face and as cross-entropy (CE), Kullback-Leibler (KL) divergence, and intersection of histograms of
  - normalized pixel luminance values of the left and right halves of the face and
  - LBP (local binary pattern) filtered pixel luminance values of the left and right halves of the face, respectively,
- characteristics of the capture process: contrast, global contrast factor, measures of image brightness (mean, variance, skewness, and kurtosis of pixel luminance values), exposure, sharpness, inter-eye distance, and blur [ISO10].

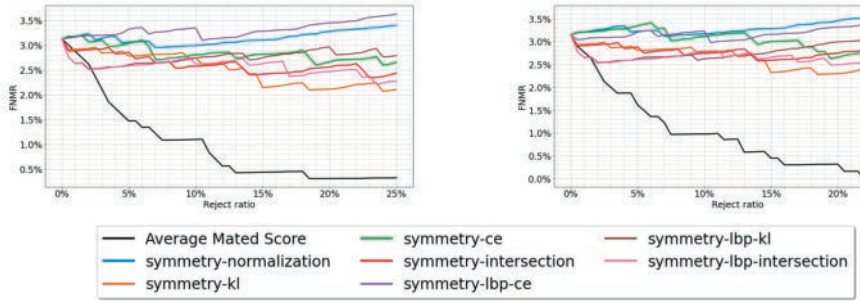
Tab. 1 shows the coefficients of the Spearman's rank correlations between the face image features and utility scores for System 1 and System 2.

Tab. 1: Spearman's rank correlation coefficients for the face image features under consideration

	symmetry-normalization	symmetry-KL	symmetry-CE	symmetry-intersection	symmetry-LBP-CE	symmetry-LBP-KL	symmetry-LBP-intersection	contrast	global contrast factor	mean of luminance	variance of luminance	skewness of luminance	kurtosis of luminance	exposure	sharpness	inter-eye distance	blur
System 1 utility score	4	-21	24	14	-8	-5	7	24	-24	27	19	-25	-26	32	27	-2	25
System 2 utility score	5	-14	27	7	-2	-10	12	21	-8	20	17	-12	-14	28	24	11	6

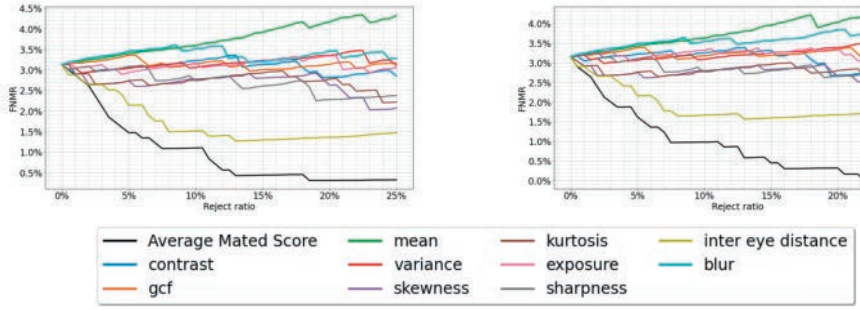
Within the training dataset, higher correlations with the utility scores were observed for exposure, mean, variance, skewness, and kurtosis of pixel luminance, left-right symmetry calculated as cross-entropy of histograms of normalized pixel values, sharpness, blur, global contrast factor, and contrast. Variance of pixel luminance was strongly correlated with contrast. Skewness of pixel luminance was strongly correlated with kurtosis. Therefore, the variance and skewness of pixel luminance need not be used in the training process.

Fig. 1 shows error vs. reject curves (ERCs) for the symmetry features, Fig. 2 for the other features. An ERC shows the dependence of the FNMR at a fixed decision threshold on



(a) System 1 (b) System 2

Fig. 1: Error vs. reject curves for symmetry features



(a) System 1 (b) System 2

Fig. 2: Error vs. reject curves for other face image features

the percentage of reference and probe images excluded based on unfavourable feature values [GT07]. The ERCs vary for different decision threshold values. The thresholds were set to give an initial FNMR value of about 3%. The ERCs show that exclusion based on the values of the features with higher correlation with utility led to reduced FNMR values within the training dataset. In addition, exclusion of images based on inter-eye distance, left-right symmetry calculated as Kullback-Leibler divergence of histograms of normalized pixel values, histogram intersection of normalized pixel values, and histogram intersection of LBP filtered pixel values led to reduced FNMR values within the training dataset. Left-right symmetry calculated as histogram intersection of normalized pixel values was strongly correlated with that calculated as Kullback-Leibler divergence of histograms of normalized pixel values and, therefore, need not be used in the training process.

From the above evaluations, we selected the following features for use in the training:

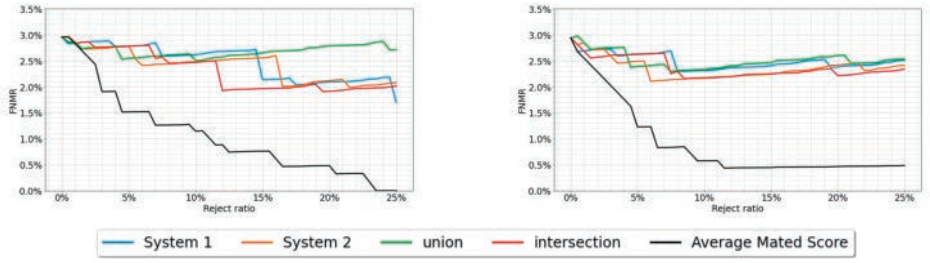
- left-right symmetry calculated as cross-entropy of histograms of normalized pixel values, as Kullback-Leibler divergence of histograms of normalized pixel values, and as histogram intersection of LBP filtered pixel values,
- from the characteristics of the capture process: contrast, global contrast factor, mean and kurtosis of pixel luminance, exposure, sharpness, inter-eye distance, and blur (i.e. all except of variance and skewness of pixel luminance).

4.2 Building utility-prediction models

For predicting the utility of face images within System 1 and System 2 and in general, random decision forests were built in Python using the training dataset. To find optimal parameter settings for the random forests, a grid search was applied, and all possible parameter combinations within the search space were verified with 3-fold cross-validation.

5 Evaluation of the accuracy of the utility-prediction model

To evaluate the accuracy of the utility-prediction models, the models trained for System 1 and System 2 and the models built using the union and the intersection of the images selected for System 1 and System 2 were used to generate quality scores for the testing data. Fig. 3 shows the ERCs with respect to these quality scores, starting at an FNMR value of about 3%. The ERCs show that exclusion of images with low quality scores lead to a reduced FNMR within the testing dataset. The model built using the intersections of images provided better results than the model built using their union did.



(a) System 1 (b) System 2

Fig. 3: Error vs. reject curves for quality scores

6 Summary and outlook

This study provided preliminary insights into features usable to predict the utility of face images within face recognition systems. Future work will expand the range of features and face recognition systems and the size and diversity of datasets explored. A next step is the evaluation of features expressing the degree of ICAO compliance. Another step is the consideration of a face image dataset containing images of all quality levels, including marginal and unacceptable quality.

Acknowledgments

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

References

[C+09] S. Curry et al. NIST Special Database 32 – Multiple Encounter Dataset I (MEDS-I) – Data Description Document. NIST Interagency Report 7679, NIST, 2009.

[F+12] M. Ferrara et al. Face Image Conformance to ISO / ICAO Standards in Machine Readable Travel Documents. *IEEE Trans. Inf. Forensics Secur.*, 7(4):1204–1213, 2012.

[G+20] P. Grother et al. Ongoing Face Recognition Vendor Test (FRVT) – Part 5: Face Image Quality Assessment. Draft NIST Interagency Report, NIST, 2020. Retrieved from https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.

[GLZ07] X. Gao, S.Z. Li, and P. Zhang. Standardization of face image sample quality. In S.-W. Lee and S.Z. Li, editors, *Proc. of ICB*, 2007.

[GNH18a] P. Grother, M. Ngan, and K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) – Part 1: Verification. Draft NIST Interagency Report, NIST, 2018. Retrieved from https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing.

[GNH18b] P. Grother, M. Ngan, and K. Hanaoka. Ongoing Face Recognition Vendor Test (FRVT) – Part 2: Identification. NIST Interagency Report 8238, NIST, 2018.

[GT07] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):531–543, April 2007.

[HO+19] J. Hernandez-Ortega et al. FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In *Int. Conf. on Biometrics (ICB)*, 2019.

[ICA18] Portrait quality (reference facial images for MRTD). ICAO Technical Report, 2018.

[ISO10] Information technology – Biometric sample quality – Part 5: Face image data. Technical Report ISO/IEC TR 29794-5, 2010.

[ISO15] Information technology – Biometric sample quality – Part 6: Iris image data. International Standard ISO/IEC 29794-6, 2015.

[ISO16] Information technology – Biometric sample quality – Part 1: Framework. International Standard ISO/IEC 29794-1, 2016.

[ISO17] Information technology – Biometric sample quality – Part 4: Finger image data. International Standard ISO/IEC 29794-4, 2017.

[ISO18] Information technology – Biometric application programming interface – Part 1: BioAPI specification. International Standard ISO/IEC 19784-1, 2018.

[ISO19] Information technology – Extensible biometric data interchange formats – Part 1: Framework. International Standard ISO/IEC 39794-1, 2019.

[T+16] E. Tabassi et al. NFIQ 2.0 – NIST Fingerprint Image Quality. Draft NIST Interagency Report, NIST, 2016. Retrieved from https://www.nist.gov/document/nfiq2reportpdf.

[TGS11] E. Tabassi, P. Grother, and W. Salamon. Performance of iris image quality assessment algorithms. NIST Interagency Report 7820, NIST, 2011.

[TWW04] E. Tabassi, C.L. Wilson, and C.I. Watson. Fingerprint image quality. NIST Interagency Report 7151, NIST, 2004.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 281

Toward to Reduction of Bias for Gender and Ethnicity from Face Images using Automated Skin Tone Classification

David Molina[1], Leonardo Causa[2], Juan Tapia[3]

Abstract: This paper proposes and analyzes a new approach for reducing the bias in gender caused by skin tone from faces based on transfer learning with fine-tuning. The categorization of the ethnicity was developed based on an objective method instead of a subjective Fitzpatrick scale. A K-means method was used to categorize the color faces using clusters of RGB pixel values. Also, a new database was collected from the internet and will be available upon request. Our method outperforms the state of the art and reduces the gender classification bias using the skin-type categorization. The best results were achieved with VGGNET architecture with 96.71% accuracy and 3.29% error rate.

Keywords: Gender classification, Bias, Skin-Detection.

1 Introduction

Facial recognition is the process of identifying or verifying the identity of a person using their face. It uses biometric to capture, analyze, and compare patterns based on the person's facial details. Traditionally, facial recognition has been associated with the security sector[BJ18]. However, it is now expanding across many other applications. Although positive results have been achieved in this field. There are still variables that limit the practical and cross-wise application of this technology, such as the accuracy and the throughput speed.

About the accuracy, some biases have been identified, mainly in the soft biometric features: gender, race and age [BJ18, JTF19, SWD19]. Several researchers have identified these types of biases, but there is not yet evidence about the real causes of such problems [BJ18]. Considering the rapid growth in the use of these technologies, the results of facial recognition systems must not be determined by some kind of algorithmic discrimination, i.e., producing better or worse results with certain groups of people, given their gender, race or age. Different studies show that facial recognition systems have higher error rates in people with dark skin, female and young [KVK12]. These shortcomings are due, in part, to biased training processes; which is facilitated by the use of databases that do not take into account the human particularities and differences, especially in aspects such as race and gender [JH14].

Several research groups have worked on facial recognition and the associated biases. Buolamwini and Gebru [BJ18] focus on evaluating the performance of three commercial

[1] Universidad Andres Bello, DCI, Avenida Antonio Varas 880, Santiago Chile, d.molinagarrido@uandresbello.edu
[2] TOC Biometrics, R+D Center SR-226, leonardo.causa@toc.cl
[3] Corresponding author: Universidad de Santiago de Chile, Departamento de Informatica, juan.tapia.f@usach.cl

classifiers, IBM [IBM20], Microsoft [MIC20], and Face++ [Fmm20]. Previous research showed discrimination according to race and gender in machine learning algorithms. The analysis of facial databases such as IJB-A [ijb-a], revealed an over-representation of lighter-skinned, compared to darker-skinned individuals, especially female. To test the classifiers, an annotated database with 1,270 images was generated, The Pilot Parliaments Benchmark (PPB) [BJ18]. The images were selected from three African countries and three European countries. They were manually grouped by gender and skin type labeled using the Fitzpatrick scale [FTZ8], and the intersection of gender and skin type. Test results showed a relatively high accuracy overall. However, the error rates increase between the different groups. All classifiers provided the best results on the males than females with an error rate between 8.1%-20.6%. Similarly, classifiers showed better performance on lighter-skinned than darker-skinned individuals, with an error rate of 11.8%-19.2%. The best results are for lighter males with 100% accuracy; while the highest error rate was for darker females ranging from 20.8% to 34.7%.

Muthukumar et al. [MPR18] conducted several analyses to try to uncover the reason for the unequal performance of commercial facial recognition services in the gender classification task across intersectional groups defined by skin type and gender. In this study, a modified PPB database [BJ18] was used: labels related to skin tone were classified only as light or dark and 1,204 images were used (PPB*). To perform tests on this data set, the IBM Watson classifier and a custom classifier were applied. Both systems showed better results on males than females and the highest error rate was for darker females ranging from 17.0% to 27.0%. The main finding is that the skin type is not the cause of misclassification. Besides, they have shown evidence suggesting differences in the lip, eyes and cheek structure by the ethnicity.

Wu and Wang [SWD19] used deep learning method to classify facial features and to study the factors affecting face recognition, mainly the influence of the age and gender. Their results showed an average recognition rate of 83.7% using the CAS-PEAL face database [CP04] (12,000 images). About the gender, the system performs better on males than females. Considering the age, middle-aged men presented lower performance than that of youth and the elderly; and the female had not a significant difference in the recognition rate. Dhomne et al. used Deep Convolutional Neural Network (D-CNN) algorithm based on a VGGNET architecture [SZ15] to develop a gender classification system. A gender-balanced database consisting of 200 celebrity images was used. Their results achieved 95.0% accuracy in the test dataset. Borza et al. [BD18] compared two methods to develop an automated skin tone classification system to use in visage applications. The first method used histograms in various color spaces and Principal Component Analysis to generate a feature vector. Afterward, a Support Vector Machine (SVM) and voting schema are used to determine the skin tone. The second method uses Convolutional Neural Networks (CNN) to automatically extract chromatic features. Both methods were trained and tested on publicly available datasets with 9,951 images: Caltech, the Chicago face dataset, Minear-Park, and Brazilian face dataset. The SVM method showed an accuracy of 86.7%, and the CNN approach obtained an accuracy of 91.3%.

The relates work shows that the main problems in gender classification is a non-representative database and manual skin type classification by human experts are critical problems in this

field of research. A balanced database in terms of race and gender and automated detection of skin tone could be a powerful tool to reduce biased, subjectivity, standardize criteria, and to improve the gender classification in facial recognition systems.

The goal of this paper is to develop a method for gender classification with racial analysis using automated detection of skin tone based on machine learning and deep learning algorithms. Additionally, we build an annotated gender and skin-type balanced-database to train and test this work. The database will be available to other researchers upon request.

2 Methods

2.1 Images Database

In order to study the bias of gender and ethnicity because of the subjective method used to label the skin-color, a new database was created collected images from the internet. This database is distributed equally in gender, which provides phenotype and geographical differentiation (see Fig. 1).

The database consists of 12,000 facial images of different phenotype groups. Divided into a Set 1 of dark-skinned people (black race) and a Set 2 of Asian and Caucasian people (white race).

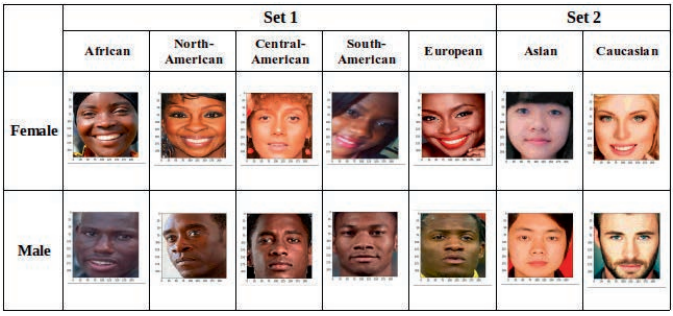


Fig. 1: Example of our annotated gender and skin-type balanced-database. Source: Self-production.

Set 1 is formed by 2,000 images of Africans, 2,000 images of African-Americans (1,000 North-Americans, 500 Central-Americans and 500 South-Americans), and 2,000 images of Europeans. The images were obtained from different existing facial databases and supplemented by Google images.

Set 2 is formed by 3,000 images of Asians and 3,000 images of Caucasians. The images were obtained from UTKFace and SCUT-FBP5500 databases. The gender in both sets is represented by 50% men and 50% female. Images dimensions are at least 250×250 pixels, a maximum of five images per subject in different positions, and the wild pose (without restrictions) were used.

2.2 Gender Classification System with Racial Analysis

The proposed classification method consists of two modules. Module 1 applies advanced image processing tools and machine learning to automatically classify the skin tone. It can be described as an analysis cascade of three stages: in Stage 1, the images are processed using CNN algorithms for face detection. Stage 2 uses skin segmentation based in HSV color space thresholds on face images to obtain the face skin [BD18][ZSQ9]. Stage 3 applies K-means [MQC7] to determine the predominant color in face skin. Module 2 uses features extraction and two classifiers D-CNN, based on VGGNET [SZ15] and MobileNet [HZ17] architectures, to generate the gender classification system.

2.3 Automated Skin Tone Classification

2.3.1 Face Detection

Stage 1 uses the CNN pre-trained algorithm based on TinyFaces detector with a ResNet-101 architecture to identify faces in the images of the database. This algorithm improves the detection of small objects [HR17] and performs well on facial images with different poses and faces of different sizes.

2.3.2 Skin Segmentation

Stage 2 allows to segment the face images generated in the previous step, to obtain only the face skin (Fig. 2). HSV color space thresholds are used for this segmentation, which has been proven to give better results in skin color extraction tasks [BD18][ZSQ9].

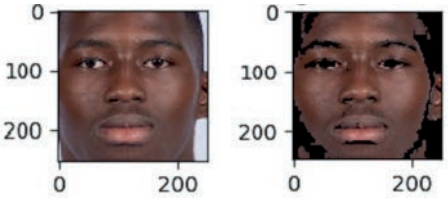


Fig. 2: Segmented image with color thresholds in the HSV color space. Source: Self-production.

2.3.3 Skin Tone Classification

The purpose of Stage 3 is to determine which is the predominant color in the skin-segmented image using the K-means algorithm. A grid search from k=2 up to 10 was used to looking at the best parameters. The best result was achieved with k=4. Each pixel on the segmented image is selected, in the RGB color space, and it is associated in one of the four clusters, depending on the chromatic differences. For each image, the most voting cluster is the predominant color cluster, i.e., it contains the type of color that is most repeated on the face, and therefore, the color with which image is classified. After all, images were associated with some clusters, the mean RGB value in Set 1, Set 2 and both are calculated to determine the thresholds of four color categories. These color categories define the predominant skin tone in the face and represent the racial analysis.

2.4 Gender Classification System

Module 2 uses deep learning to develop gender classifier. Two classification method based on D-CNN were implemented and trained: VGGNET and MobileNet. To train and test, the image database was divided into three different sets: training set, validation set and testing dataset. The output of Module 2 and the skin tone are used to analyze and evaluate the system in terms of gender, skin-type and intersectional groups.

3 Results

3.1 Automated Skin Tone Classification

Tables 1 and 2 shows the distribution of the RGB component (mean and standard deviation) for Set 1 and Set 2 by geographical zones and gender.

Zone	**Set 1 - Male**		**Set 1- Female**	
	Mean RGB	**SD RGB**	**Mean RGB**	**SD RGB**
African	84.25	27.36	90.90	31.68
South-American	96.43	35.03	111.32	40.73
Central-American	98.29	34.47	110.45	41.23
North-American	95.28	27.51	107.81	30.66
European	90.95	29.30	113.14	40.55

Tab. 1: RGB Component for Set 1.

Zone	**Set 2 - Male**		**Set 2- Female**	
	Mean RGB	**SD RGB**	**Mean RGB**	**SD RGB**
Asian	160.28	34.13	172.53	33.44
Caucasian	150.17	37.03	161.61	39.40

Tab. 2: RGB Component for Set 2.

To define the category thresholds, the mean RGB value of each set and the total database was used. Categories 1 and 2 represent dark skin tones. Categories 3 and 4 represent light skin tones. The categories 1, 2, 3 and 4 reached the following skin tone values respectively: $<=$ 97.48, [97.48 - 129.32], [129.32 - 161.15] and $>=$ 161.15.

3.2 Gender Classification System

The system was trained using a person-disjoint database and the parameters were adjusted using three partition sets: Train, Validation and Test. A training set of 60% (7,200) and validation set of 20% (2,400) was used. The performance of the all system was measured using the testing set of 20% (2,400) dataset. For both models, different D-CNN configurations and parameters tuning were applied. The best results were obtained for the models with data augmentation, 150 epochs, inputs image size 224×224 pixels and learning rate of $1e-4$. Some results are shown below.

3.2.1 VGG16 Net Architecture

The overall results for gender classification show a 96.71% accuracy and 3.29% error rate (4.17% Set 1 and 2.42% Set2 -lighter-skinned group). Table 3 shows the error by skin tone category and gender. In Tables 4 and 5 the error classification rate is showed by gender and geographical zone for each data set. In all tests, the error increased for the darker group (Set 1 and category 1) compared with the lighter group (Set 2 and category 4). The highest error rate was for the darker female group.

Category	Female		Male	
	Amount	Error [%]	Amount	Error [%]
1	22	5.76	15	3.92
2	9	4.52	8	4.02
3	7	3.48	3	1.50
4	10	2.40	5	1.20
Total	48	4.00	31	2.58

Tab. 3: Classification Error rate by Skin Tone Categories and Gender.

Zone	Set 1 - Female		Set 1 - Male	
	Amount	Error [%]	Amount	Error [%]
African	13	6.50	12	6.00
South-American	2	4.00	1	2.00
Central-American	2	4.00	2	4.00
North-American	6	6.00	2	2.00
European	7	3.50	3	1.50
Total	30	5.00	20	3.33

Tab. 4: Classification Error Rate by Gender and Geographical Zone on Data Set 1. ”Amount” represents the number of images miss-classified.

Zone	Set 2 - Female		Set 2 - Male	
	Amount	Error [%]	Amount	Error [%]
Asian	10	3.33	6	2.00
Caucasian	8	2.67	5	1.67
Total	18	3.00	11	1.83

Tab. 5: Classification Error rate by Gender and Geographical Zone on Data Set 2.

3.2.2 MobileNet Architecture

The overall results for gender classification show a 96.33% accuracy and a 3.67% error rate (4.17% Set 1 and 3.17% Set 2-lighter-skinned group) but with a small increase in error rate in this set.

In Tables 6 and 7 the error classification rate is showed by gender and geographical zone for each data set. Table 8 shows the error by skin tone category and gender. The error increased for the darker group compared with the lighter group. Unlike the previous case, there is an improvement in female classification, being the highest error rate for males, especially for the darker male group.

Zone	Set 1 - Female		Set 1 - Male	
	Amount	Error [%]	Amount	Error [%]
African	11	5.50	13	6.50
South-American	1	2.00	2	4.00
Central-American	1	2.00	3	6.00
North-American	3	3.00	1	1.00
European	7	3.50	8	4.00
Total	23	3.83	27	4.50

Tab. 6: Classification Error rate by Gender and Geographical Zone on Data Set 1.

Zone	Set 2 - Female		Set 2 - Male	
	Amount	Error [%]	Amount	Error [%]
Asian	4	1.33	15	5.00
Caucasian	8	2.67	11	3.67
Total	12	2.0	26	4.33

Tab. 7: Classification Error rate by Gender and Geographical Zone on Data Set 2.

Category	Female		Male	
	Amount	Error [%]	Amount	Error [%]
1	15	3.93	21	5.48
2	12	6.00	7	3.50
3	5	2.49	8	4.00
4	3	0.72	14	3.36
Total	35	2.92	50	4.17

Tab. 8: Classification Error Rate by Skin Tone Categories and Gender.

3.3 Comparison with Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Gender shades [BJ18] is one of most relevant work about gender and ethnicity bias. A manually Fitzpatrick skin-type scale [FTZ8] was used to labeling face images (PPB database) in six categories (Types I to VI). Faces labeled were grouped in two skin tone groups, lighter skin (Types I, II and III) and darker skin (Types IV, V, and VI). This classification reached 46.4% for darker skin and 53,6% for lighter skin. Our proposal clustering automatically the same four categories using K-means as reported in [HR17], a lighter skin that includes skin-tone categories 3 and 4, and a darker skin that includes skin-tone categories 1 and 2.

Table 9a presents the distribution of both databases, showing a similar proportion of skin types. The overall accuracy is presented in Table 9b. Our model shows the best results with an improvement ranged from 3.0% up to 8.8%.

In terms of gender (Table 10a) and skin type distribution (Table 10b), our results are better than test reported for the commercial classifiers [GS18]. Gender classification shows an error rate difference of 1.42% compared with the 8.1% for Microsoft, which was obtained

Skin Type	Our Work [%]	Gender Shades [%]
Darker Skin	48.5	46.4
Lighter Skin	51.1	53.6

(a)

Classifier	Accuracy [%]
Our Work	96.7
Microsoft	93.7
Face++	90.0
IBM	87.9

(b)

Tab. 9: a) Database Distribution by Skin Type. b) Overall Accuracy.

the lowest error rate difference between commercial classifiers. Considering skin-type, the error rate was of 2.5%, while the best results for commercial classifiers were obtained by Face++ with an 11.8% error rate. Table 11 shows the results by intersectional groups perform worst on darker females, but our method presents a great improvement by reducing the gap to 3.6%.

Classifier	Female[%]	Male[%]	Error[%]
Our work	96.0	97.4	1.4
Microsoft	89.3	97.4	8.1
Face ++	78.7	99.3	20.6
IBM	79.7	94.4	14.7

(a)

Classifier	Darker [%]	Lighter[%]	Error[%]
Our Work	95.4	97.9	2.5
Microsoft	87.1	99.3	12.2
Face++	83.5	95.3	11.8
IBM	77.6	96.8	19.2

(b)

Tab. 10: a) Accuracy by gender. b) Accuracy by skin type.

Classifier	**DM [%]**	**DF [%]**	**LM [%]**	**LF [%]**	**Gap [%]**
Our Work	95.6	95.2	96.7	98.8	3.6
Microsoft	94.0	79.2	100.0	98.3	20.8
Face++	99.3	65.5	99.2	94.0	33.8
IBM	88.0	65.3	99.7	92.9	34.4

Tab. 11: Overall Accuracy by Gender and Skin Type: Darker Male (DM), Darker Female (DF), Lighter Male (LM) and Lighter Female (LF).

4 Conclusion

In this ongoing research, we show that is feasible to develop an objective method to assign the skin-tone in order to improve the gender classification. This approach improves the results and reduces gender bias by ethnicity produced by the manual assignation of each categorization. This assignation is influenced by the experience of each people. Another achievement of this work is the construction of an annotated gender and skin-type balanced-database, which can be used to train and test this and other methods upon request.

References

[BG18] J. Buolamwini and T. Gebru, “Gender shades: intersectional accuracy disparities in commercial gender classification”,Proceeding of Machine Learning Research, vol. 81, pp. 1—15, 2018.

[TP19] Juan Tapia and Claudio Perez, “Clusters of Features using Complementary Information Applied to Gender Classification From Face Images,” in IEEE Access, vol. 7, pp. 79374-79387, 2019. doi: 10.1109/ACCESS.2019.2923626.

[WW19] S. Wu and D. Wang, “Efect of subjects age and gender on face recognition results,” Journal of Visual Communication and Image Representation, vol. 60, pp.116–122, 2019.

[KBK12] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, “Face recognition performance: Role of demographic information, ” IEEE Transactions on Information Forensics and Security, vol. 7(6), pp. 1789–1801, 2012.

[JH14] A. K. Jain Hu Han, “Age, gender and race estimation from unconstrained face images,” MSU Technical Report: MSU-CSE-14-5, 2014.

[Ibm20] IBM, “Watson Visual Recognition - Overview,” 2020. [Online]. Available: https://www.ibm.com/cloud/watson-visual-recognition. [Acceded: Feb. 10, 2020].

[Mic20] Microsoft Inc., “Cognitive Services - APIs for AI Developers,” 2020. [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/. [Acceded: Feb. 10, 2020].

[Fa20] Face++, “Face Detection - Cognitive Services,” 2012-2020. [Online]. Available: https://www.faceplusplus.com/face-detection/. [Acceded: Feb. 10, 2020].

[KKT] B. F. Klare, B. Klein, E. Taborsky, A Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, “Pushing the frontiers of unconstrained f.ace detection and recognition: Iarpa janus benchmark,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1931–1939, 2015.

[Fit8] T. B. Fitzpatrick, “The Validity and practicality of sun-reactive skin types I through VI,” Archives of Dermatology, vol. 124(6), pp. 869, 1988.

[MPR18] V. Muthukumar, T. Pedapati, N. Ratha, P. Sattigeri, C-W. Wu, B. Kingsbury, A. Kumar, S. Thomas, A. Mojsilovic, and K. R. Varshney, “Understanding unequal gender classification accuracy from face images,” 2018.

[GCS04] W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, “The CAS-PEAL large-scale Chinese face database and baseline evaluations,” ICT-ISVISION Joint Research & Development Laboratory for Face Recognition, Chinese, 2004.

[SZ15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR, 2015.

[BDD18] D. Borza, A. Darabant, and R. Danescu, “Automatic skin tone extraction for visagism applications,” In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. SCITEPRESS - Science and Technology Publications, 2018.

[ZSQ9] B. D. Zarit, B. J. Super, and F. Quek, “Comparison of five color models in skin pixel classification,” Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV. IEEE Comput. Soc, 1999.

[Mac7] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” Proc. of rhe Fifh Berkeley Symposium on Math., Srar. and Prob., Vol. 1, pp. 281-296, 1967.

[HZ17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017.

[HR17] P. Hu and D. Ramanan, “Finding tiny faces,” In 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2020 291

Compact Models for Periocular Verification Through Knowledge Distillation

Fadi Boutros[12], Naser Damer[12], Meiling Fang[12], Kiran Raja[3], Florian Kirchbuchner[1], Arjan Kuijper[12]

Abstract: Despite the wide use of deep neural network for periocular verification, achieving smaller deep learning models with high performance that can be deployed on low computational powered devices remains a challenge. In term of computation cost, we present in this paper a lightweight deep learning model with only 1.1m of trainable parameters, DenseNet-20, based on DenseNet architecture. Further, we present an approach to enhance the verification performance of DenseNet-20 via knowledge distillation. With the experiments on VISPI dataset captured with two different smartphones, iPhone and Nokia, we show that introducing knowledge distillation to DenseNet-20 training phase outperforms the same model trained without knowledge distillation where the Equal Error Rate (EER) reduces from 8.36% to 4.56% EER on iPhone data, from 5.33% to 4.64% EER on Nokia data, and from 20.98% to 15.54% EER on cross-smartphone data.

Keywords: Periocular recognition, Smartphone biometric verification, Knowledge distillation.

1 Introduction

The rapid growth of smartphone users (3.2 billion in 2019 [St20]) has also increased the interest in secure authentication application using smartphones. Biometric modalities like fingerprint, voice, periocular and face are widely employed on smartphones to achieve secure, convenient, and reliable authentication.

Of the many other modalities, periocular region provides a distinct trade-off between using iris or entire face for identity verification by considering a small area around the eye including eyelids, lashes, and eyebrows as biometric trait [PRJ09]. Given the performance under relaxed settings, periocular biometrics is recently well preferred for various use cases such as mobile platform [Al19] and embedded device [Bo19, Bo20a, Bo20b]. Motivated by such new applications, we focus on periocular modality for smartphone based biometric identity verification in this work.

Although the integration of biometrics in smartphone devices has enabled several advantages, deploying such a solution to a smartphone device faces several challenges. One of these challenges is the high variability between probe and gallery images produced when the images are acquired using different devices, different cameras, or under different environmental conditions, requiring a highly generalized solution. This challenge is well addressed in the literature as reported in the previous works [Al19, Ah17]. Yet another major challenge is related to the limited computational resources available in smartphone

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany
[3] The Norwegian Colour and Visual Computing Laboratory, NTNU, Gjovik, Norway

devices, especially when considering a solution based on a deep neural network with extremely high number of parameters. Recent works [Ga18, Ah17] have addressed this issue with the smartphone based periocular recognition using deep learning, albeit with less focus on the limited computation resource available on the smartphone devices where both models [Ga18, Ah17] contain more than 12 million of trainable parameters. Despite the use of deep learning, the challenge of customizing the solutions to smartphone devices with limited computational resources is not well addressed.

We therefore focus this work on reducing the number of parameters in the deep models to make them easily adaptable to mobile devices with limited computation resources by utilizing knowledge distillation noted as KD [HVD15] for periocular verification. To truly establish the applicability of the proposed approach for periocular verification, we provide the baseline performance of three DenseNet architectures [Hu17]: DenseNet-201, DenseNet-169, and DenseNet-121. Further, we propose a compact model which we refer to as DenseNet-20 based on the dense block containing 1.1 million of trainable parameters. The experimental results on VISPI dataset [KRB20] of 152 unique pericoular instances with 6682 images captured with 2 different smartphones (iPhone 5S and Nokia Lumia 1020) shows that the DenseNet-20 model achieves a comparable verification performance using a shallow architecture. With the obtained performance, we argue that deploying such a model to a low computational resource device is more realistic than other deeper models. Motivated by this, we also focus on enhancing the accuracy and generalizability of the shallow model for periocular recognition by successfully introducing the KD method to the training process. Although introducing knowledge distillation to the training process does not change the model capacity, the gradient descent induced by distillation loss function allows this model to find a very favorable minimum of the training objective [PL19]. Thus, our proposed approach improves the verification performance of the distilled model, in comparison to the same model trained without knowledge distillation, the Equal Error Rate (EER) is reduced from 8.36% to 4.56% on iPhone data, from 5.33% to 4.64% EER on Nokia smartphone data, and from 20.98% to 15.54% EER on cross-smartphone data.

2 Methodology

The goal of this work is to present a solution to improve the accuracy and generalizability of shallow CNN models for smartphone periocular verification. Particularly, we first evaluate deep representations extracted from periocular region using three different DenseNet [Hu17] architectures: DensNet-121, DensNet-169, and DenseNet-201. We further present our proposed compact CNN model, DenseNet-20, containing only 1.1 million trainable parameters. To further improve the generalizability and accuracy of the small CNN model, we introduce knowledge distillation (KD) to the DenseNet-20 model training process. This section presents the details of the employed DenseNet model along with the KD method.

2.1 Densely Connected Convolutional Networks

DenseNet [Hu17] is a convolutional neural network designed for image classification to achieve low classification error rates while having fewer parameters than ILSVRC 2015 winner, ResNet model [He16]. The architecture is based on connecting each convolutional layer to every other layer in a feed-forward fashion as shown in Figure 1. Thus, each layer

ℓ^{th} receives collective knowledge from all preceding layers $x_0, x_1, ..., x_{\ell-1}$ and passes on its knowledge to all subsequent layers. Given that each layer produces k feature maps, the input feature map for ℓ^{th} layer is $k_0 + k \times (\ell - 1)$ where k_0 is the number of channels in the input layer and k refers to the growth rate of the network. In this work, we evaluate three different DenseNet architectures as baselines: DenseNet-121, DenseNet-169, and DenseNet-201 where 121, 169, and 201 refer to the number of the convolutional layers in each model (network depth). The growth rate for all the networks is set to $k = 32$. The DenseNet-121, DenseNet-169, and DenseNet-201 models contain 7.1, 12.6 and 18.2m of trainable parameters, respectively.

We apply transfer learning on these models pretrained on ImageNet dataset [De09] by fine-tuning all the layers on images from our training dataset with Softmax classifier. In the test phase, the Softmax classifier is removed from all models and the feature f is extracted from the last layer which is of the dimension $7 \times 7 \times 1920$.

2.2 Proposed Compact DenseNet

We further propose a new model based on DenseNet architecture - DenseNet-20. Similar to the original DenseNet model, DenseNet-20 has 4 dense blocks with 1, 2, 8, and 6 layers in dense block 1, 2, 3, and 4, respectively. We train the compact DenseNet-20 model from scratch with Softmax classifier. The proposed DenseNet-20 contains 1.1m trainable parameters as compared to 18.2 million parameters with DenseNet-201. Similar to the original DenseNet models, the Softmax classifier is removed in the testing phase from the model to extract the feature f from the last layer with the dimension of $7 \times 7 \times 1920$.

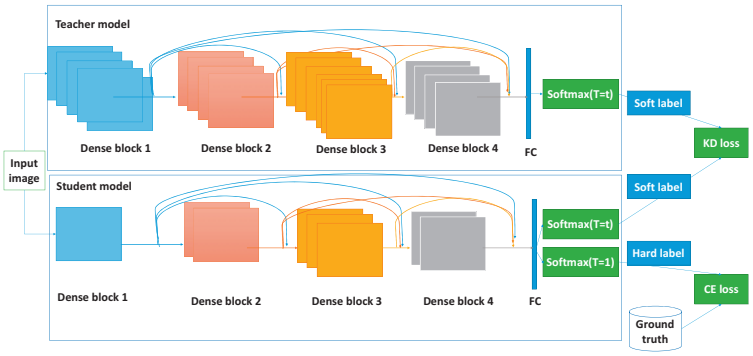


Fig. 1: An overview of the proposed KD approach for periocular verification based on DenseNet architecture.

2.3 Proposed Compact DenseNet-20 with Knowledge Distillation

We explore KD to improve the performance of DenseNet-20 model by employing a student-teacher relation where each of DenseNet-121, DenseNet-169, and DenseNet-201 models are used as a teacher to distill the knowledge to student model, DenseNet-20. We present the details of KD for the convenience of the readers.

KD is a technique to improve the performance and generalization ability of smaller models by transferring the knowledge learned by a cumbersome model (teacher) to a single

small model (student). The key idea is to guide the student model to learn the relationship between different classes discovered by the teacher model that contains more information beyond the ground truth labels [HVD15]. Suppose we have teacher model T, student model S, and training dataset $X, Y \in D$, where X is the training images and Y is their class labels. The output of the teacher model for any input $x_i \in X$ is a vector of class probabilities P^T computed for each class using softmax function by converting the logits, z^T into probabilities that sum to one $P^T(x) = softmax(z^T)$. Specifically, the probability p_i of class i is computed by comparing z_i with other logits as given: $p_i = \frac{exp(z_i)}{\sum_{j=1}^{N} exp(z_j)}$. This probability distribution will have a high probability value of p_i for the correct class $y_i \in Y$ with all other class probabilities close to zeros. Thus, it does not provide more valuable information than ground truth labels. Therefore, Hinton et al. [HVD15] proposed to scale the logits using a temperature parameter $t > 1$ before applying the softmax function. Thus, the teacher model can produce a softer distribution of the class probabilities, which provides more valuable information about classes similar to the predicted class. In this case, the output of the teacher model is $P_s^T(x) = Softmax(z^T/t)$ and the probability p_i of class i is given as: $p_i = \frac{exp(z_i/t)}{\sum_{j=1}^{N} exp(z_j/t)}$. Similarly, student S can produce a soft class probability distribution using the temperature parameter t, $P_s^S(x) = Softmax(z^S/t)$. The final loss for the student model is a weighted sum of two loss functions, cross-entropy loss L_{ce} and Kullback Leibler Divergence loss L_{kld}, as follows:

$$L_{KD} = \lambda * L_{ce}(Y, P^S(x)) + (1-\lambda) * t^2 * L_{kld}(P_s^S(x), P_s^T(x)),$$

where Y is the ground truth label, $P^S(x)$ standard softmax output produced by student, $P_s^S(x)$ parameterized softmax output produced by student, $P_s^T(x)$ parameterized Softmax output produced by teacher and $\lambda \in [0,1]$ is the weight parameter. Since the gradients of the L_{ce} loss is smaller than gradients of the L_{kld} where the logits used for L_{kld} is divided by t, the L_{kld} is multiplied by t^2 as suggested by Hinton et al. [HVD15].

We thus use the student-teacher based KD for all three DenseNet models - DenseNet-201, DenseNet-169 and DenseNet-121 by setting each of them as teacher and our proposed DenseNet-20 as the student model as shown in the Figure 1.

3 Experimental setup

To demonstrate the applicability of our proposed approach, we evaluate it on a public dataset of periocular images - VISPI database [KRB20]. We employ the subset of database containing 152 unique periocular instances captured from 76 unique subjects using two different smartphones - iPhone 5S and Nokia Lumia 1020. The 152 periocular instances are captured from both left and right eyes- 76 instances are captured from the left eye and 76 instances are captured from the right eye. Each unique periocular image has multiple samples captured in different instances. The total distribution of the images in the database used for the evaluation in this work is presented in the Table 1.

Details	Smartphone	
	iPhone 5S	Nokia Lumia 1020
Capture Scenario	Mixed Illumination	Mixed Illumination
Resolution	12 Mp	41 Mp
Number of subjects	76	76
Unique periocular instances	152	152
Total images	3341	3341

Tab. 1: Distribution of periocular database employed in this work.

The ocular images are captured in a mixed illumination environment using the rear camera of the smartphones in a semi-cooperative manner. The images in the database also present everyday appearance variations that include the make-up and non-uniform illumination. Beside, the images in the VISPI database present various forms of degradation due to motion blur and eye blinking. Further, the influence of both the external sunlight illumination and the artificial room light illumination along with other degrading factors make the cross-sensor/cross-smartphone comparison challenging. The sample images from the periocular database as depicted in Figure 2 illustrate a set of variation and degradation in terms of appearance under different smartphones both across the phones and the subjects.

Of the 152 unique periocular instances, the first 100 instances (from 50 subjects, i.e., 50 instances captured from the left eye and 50 instances captured from the right eye) are used for the training and the other 52 instances (from 26 independent subjects, i.e., 26 instances captured from the left eye and 26 instances captured from the right eye) are used for testing. Further, a random subset of 200 images (two images per instance) is selected from the training split to validate the model during the training phase.

Fig. 2: Sample images from VISPI database

All the images are uniformly resized to a size of 224×224 pixels to match the input layer size of DenseNet model. The training data is augmented by applying horizontal and vertical random shifting by up to 20% of the image width and/or height, and random horizontal flipping. All models are trained with a batch size of 16 and SGD optimizer with Nesterov momentum 0.9. The initial learning rate is set to $\gamma = 0.001$ and $\gamma = 0.1$ for teacher models and student model, respectively and it is dropped by a factor of 0.1 when the accuracy on the validation dataset does not improve by a value of 0.1 for 5 consequent epochs. The initial number of epochs is set to 100 and early-stopping patience parameter is set to 10 causing DenseNet-20, DenseNet-121, DenseNet-169 and DenseNet-201 to stop after 29, 11, 11, 11 epochs, respectively. The training of the student model, DenseNet-20, trained KD loss stopped after 29, 34, and 28 epoch using teacher model DenseNet-121, DenseNet-169 and DenseNet-201, respectively. In practice, the training is performed offline once and the trained model is deployed on mobile devices, which makes the size of the model the most critical deployment factor. We followed the common choice for the KD hyperparameters [HVD15, CH19, Fu18] with Temperature $t >= 4$ and $\lambda = 0.9$.

The verification performance is reported using the cosine similarity measure for comparing the features extracted from the learnt models. The result is reported first for the DenseNet-20, DenseNet-121, DenseNet-169, and DenseNet-201 models without applying the KD. In addition, we report the result of the KD on the student model DenseNet-20 with DenseNet-121, DenseNet-169 or DenseNet-201 as a teacher which we note as as DenseNet-20-KD121, DenseNet-20-KD169 and DenseNet-20-KD201 respectively.

For each of the settings, we investigate the verification performance under three different evaluation scenarios defined as following:

- iPhone verification scenario: The reference and the probe images are acquired using the camera of iPhone smartphone.
- Nokia verification scenario: Similar to the previous scenario, the reference and the probe images are acquired using Nokia smartphone.
- Cross-smartphone verification scenario: the reference images are captured using iPhone camera and the probe images are captured using Nokia camera.

The verification performance is reported using Receiver Operating Characteristic (ROC) curves, Area under the curve (AUC), False Match Rate (FMR) at fixed False Non-Match Rate (FNMR) (FMR10, the lowest FNMR for FMR$\leq$10%), and Equal Error Rate (EER). The verification performances of the different experimental settings are presented in Figure 3 along with the EER and FMR10 values in Table 2. Each of the Figures 3.a-c shows the achieved ROC of iPhone, Nokia, and cross-smartphone verification scenario.

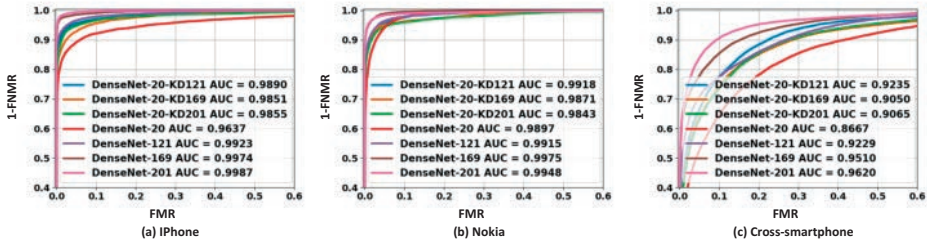


Fig. 3: The achieved ROC for different experimental settings. One can be clearly noticed the improvement in the DenseNet-20 verification performance using KD method.

4 Results and Discussion

One can clearly notice from Table 2 that the verification performances are consistently better based on the model size when same training procedure is followed. The first four rows in Table 2 present the verification performances of the DenseNet-201, DenseNet-169, DenseNet-121, and DenseNet-20 trained without the KD method. The highest verification performance among all evaluated models is achieved by the DenseNet-201 model, where the achieved EER was 9.71% for cross-smartphone verification scenario, 1.60% for iPhone verification scenario, and 2.57% for Nokia verification scenario. Also, it can be observed from the Table 2 that the DenseNet-20 model aims at maintaining (to a large degree) the verification performance of deeper model where the achieved EERs were 8.36%, 5.33% and 20.98% for iPhone, Nokia and cross-smartphone verification scenarios, respectively.

It can be further noticed that the verification performances degrade for all models when the references and probes images are captured from different smartphones in comparison to the case where the probe and the reference images are captured from the same smartphone as shown in the Table 2. However, this degradation in the performance is a common problem for cross-smartphone verification scenario as reported in the previous works [Al19].

4.1 Impact of Knowledge Distillation

The results of the proposed approach based on KD are presented in the Table 2 and Figure 3. We make the following observations from the obtained results:

Model	Inference time	Num. of. parameters	Teacher	iPhone		Nokia		Cross-smartphone	
				EER	**FMR10**	**EER**	**FMR10**	**EER**	**FMR10**
DenseNet-201	5.4ms	18.2	-	0.0160	0.0026	0.0257	0.0105	0.0971	0.0949
DenseNet-169	4.7ms	12.6	-	0.0220	0.0093	0.0256	0.0052	0.1224	0.1459
DenseNet-121	3.8ms	7.1	-	0.0396	0.0212	0.0417	0.0213	0.1666	0.2257
DenseNet-20	2.1ms	1.1m	-	0.0836	0.0782	0.0533	0.0227	0.2098	0.3556
DenseNet-20-KD201	2.1ms	1.1m	DenseNet-201	0.0515	0.0340	0.0538	0.0404	0.1709	0.2640
DenseNet-20-KD2169	2.1ms	1.1m	DenseNet-169	0.0617	0.0440	0.0496	0.0376	0.1711	0.2582
DenseNet-20-KD121	2.1ms	1.1m	DenseNet-121	**0.0456**	**0.0298**	**0.0464**	**0.0240**	**0.1554**	**0.2251**

Tab. 2: Performance obtained for different experimental settings along with inference time (in millisecond) and the number of trainable parameters (in million) for each of the evaluated models. The first four rows of the table present the achieved result for the three teacher models and for the student models (without using KD). The last three rows of the table present the achieved verification performance by including KD in the training process.

• It is noticed that introducing the KD to the DenseNet-20 model training significantly improved the verification performance and outperforms teacher model in some cases. For example, in the cross-smartphone verification scenario, the student outperformed its teacher DenseNet-121 where the achieved EER by the student was 15.54% and by its teacher was 16.66%. Similar observations is also reported in in previous work [Fu18].

• The best verification performance is achieved using DenseNet-121 model as teacher, where the achieved EERs in this case were 5.56% 4.64% and 15.54% for iPhone, Nokia and cross-smartphone verification scenarios.

• Using a larger and more accurate teacher model did not serve as better supervision to the student model as seen in Table 2. This can be explained by the fact that as the teacher model becomes more accurate using a deeper architecture, the soft probabilities produced by the teacher will contain more complex information about the class distributions and the small student model will not be able to learn all this complex information considering the small student capacity. Similar conclusion is also reported in the previous work [CH19].

5 Conclusion

We presented in this work a new approach for periocular verification exploiting the idea of Knowledge distillation (KD). The proposed models have resulted in significantly lower model size but with comparable performance to larger deep models. Through the experiments on public periocular dataset consisting of 152 unique periocular instances captured with two different smartphones, we showed that applying KD on DenseNet-20 training process achieves an EER of 4.5% on iPhone data, 4.6% on Nokia data, and 15.54% on cross-smartphone data, in comparison to EER of 8.36% on iPhone data, 5.33% on Nokia data, and 20.98% on cross-smartphone data when the same model trained without KD. In the future works in this direction, we intend to investigate the proposed method on larger datasets captured in multiple sessions to gain insights on generalizability aspects.

Acknowledgment: this research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[Ah17] Ahuja, Karan; Islam, Rahul; Barbhuiya, Ferdous A; Dey, Kuntal: Convolutional neural networks for ocular smartphone-based biometrics. Pattern Recognition Letters, 2017.

[Al19] Alonso-Fernandez, Fernando; Raja, Kiran B.; Raghavendra, Ramachandra; Busch, Christoph; Bigün, Josef; Vera-Rodríguez, Rubén; Fiérrez, Julian: Cross-Sensor Periocular Biometrics: A Comparative Benchmark including Smartphone Authentication. CoRR, abs/1902.08123, 2019.

[Bo19] Boutros, Fadi; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Eye-MMS: Miniature Multi-Scale Segmentation Network of Key Eye-Regions in Embedded Applications. In: Proceedings of the IEEE/CVF ICCVW. Oct 2019.

[Bo20a] Boutros, F.; Damer, N.; Raja, K.; Ramachandra, R.; Kirchbuchner, F.; Kuijper, A.: Periocular Biometrics in Head-Mounted Displays: A Sample Selection Approach for Better Recognition. In: 2020 8th IWBF. pp. 1–6, 2020.

[Bo20b] Boutros, Fadi; Damer, Naser; Raja, Kiran; Ramachandra, Raghavendra; Kirchbuchner, Florian; Kuijper, Arjan: Iris and Periocular Biometrics within Head Mounted Displays: Segmentation, Recognition, and Synthetic Generation. Image Vis. Comput., 2020.

[CH19] Cho, Jang Hyun; Hariharan, Bharath: On the efficacy of knowledge distillation. In: Proceedings of the IEEE ICCV. pp. 4794–4802, 2019.

[De09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Li, Fei-Fei: ImageNet: A large-scale hierarchical image database. In: CVPR 2009, 20-25 June 2009, Miami, Florida, USA. IEEE Computer Society, pp. 248–255, 2009.

[Fu18] Furlanello, Tommaso; Lipton, Zachary Chase; Tschannen, Michael; Itti, Laurent; Anandkumar, Anima: Born-Again Neural Networks. In (Dy, Jennifer G.; Krause, Andreas, eds): Proceedings of the 35th ICML 2018, Sweden. volume 80 of PMLR, pp. 1602–1611, 2018.

[Ga18] Garg, Rishabh; Baweja, Yashasvi; Ghosh, Soumyadeep; Singh, Richa; Vatsa, Mayank; Ratha, Nalini: Heterogeneity aware deep embedding for mobile periocular recognition. In: 9th BTAS. IEEE, 2018.

[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.

[Hu17] Huang, Gao; Liu, Zhuang; Van Der Maaten, Laurens; Weinberger, Kilian Q: Densely connected convolutional networks. In: Proceedings of the IEEE CVPR. pp. 4700–4708, 2017.

[HVD15] Hinton, Geoffrey E.; Vinyals, Oriol; Dean, Jeffrey: Distilling the Knowledge in a Neural Network. CoRR, abs/1503.02531, 2015.

[KRB20] Kiran Raja; Ramachandra, Raghavendra; Busch, Christoph: Collaborative representation of blur invariant deep sparse features for periocular recognition from smartphones. Image and Vision Computing, p. 103979, 2020.

[PL19] Phuong, Mary; Lampert, Christoph: Towards understanding knowledge distillation. In: International Conference on Machine Learning. pp. 5142–5151, 2019.

[PRJ09] Park, Unsang; Ross, Arun; Jain, Anil K: Periocular biometrics in the visible spectrum: A feasibility study. In: 2009 IEEE 3rd BTAS. IEEE, pp. 1–6, 2009.

[St20] Number of smartphone users worldwide from 2016 to 2021. ”https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/”.

A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 299

Can Generative Colourisation Help Face Recognition?

Pawel Drozdowski[1], Daniel Fischer[1], Christian Rathgeb[1], Julian Geissler[1], Jan Knedlik[1], Christoph Busch[1]

Abstract: Generative colourisation methods can be applied to automatically convert greyscale images to realistically looking colour images. In a face recognition system, such techniques might be employed as a pre-processing step in scenarios where either one or both face images to be compared are only available in greyscale format. In an experimental setup which reflects said scenarios, we investigate if generative colourisation can improve face sample utility and overall biometric performance of face recognition. To this end, subsets of the FERET and FRGCv2 face image databases are converted to greyscale and colourised applying two versions of the DeOldify colourisation algorithm. Face sample quality assessment is done using the FaceQnet quality estimator. Biometric performance measurements are conducted for the widely used ArcFace system with its built-in face detector and reported according to standardised metrics. Obtained results indicate that, for the tested systems, the application of generative colourisation does neither improve face image quality nor recognition performance. However, generative colourisation was found to aid face detection and subsequent feature extraction of the used face recognition system which results in a decrease of the overall false reject rate.

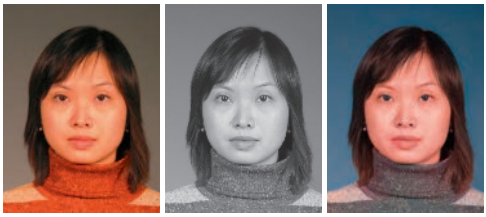
Keywords: biometrics, face recognition, face image quality, generative colourisation.

1 Introduction

Developments in deep neural networks have shown impressive improvements in diverse generative image processing tasks, *e.g.* single-image super resolution [Ha19] or inpainting [Li18]. Focusing on face images, domain-specific techniques have been established, *e.g.* face hallucination [LSF07, Ch18, GSŠ19] or face completion [Li17, Ca19]. Some of these methods have been found advantageous in various face-related vision tasks, such as face detection and recognition [Li19, MIA19]. In addition to the aforementioned generative methods, image colourisation schemes based on deep neural networks have been proposed [CYS15, ZIE16, NNE18], often for the purpose of restoring old images and film footage. Said methods are able to generate realistic colour images from greyscale images, including facial imagery. An example for applying a state-of-the-art colourisation algorithm to a face image is depicted in figure 1.

In this work, we investigate if generative colourisation can be advantageous in the context of face recognition. To this end, face image subsets of two publicly available databases [Ph98, Ph05] are converted to greyscale and colourised using two versions of a public colourisation algorithm [An19, Ke19]. Subsequently, face sample quality is assessed employing the public algorithm of FaceQnet [He19]; furthermore, standardised ISO/IEC methodology and metrics [IS06] are used to evaluate the biometric performance of the ArcFace recognition system [De19] in a scenario-based manner. The considered scenarios reflect different

[1] da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {pawel.drozdowski,daniel.fischer,christian.rathgeb,christoph.busch}@h-da.de



(a) original (b) greyscale (c) colourised

Fig. 1: Example of generative face colourisation: (a) the original image, (b) the original image converted to greyscale, and (c) the colourised greyscale image using the method of Antic [An19].

relevant use cases in which colourisation methods might be applied to reference and/or probe images prior to the feature extraction. To the best of the authors' knowledge, this reproducible study represents the first investigation of the usefulness of generative colourisation in face recognition.

The remainder of this paper is organised follows: the employed image colourisation methods are described in section 2. Relevant scenarios which are considered in experiments are summarised in section 3. Experimental setup and results are presented in section 4. Conclusions are drawn in section 5.

2 Face Image Colourisation

The goal of image colouration is the addition of colour information to greyscale images (as illustrated in figure 1). Early solutions (see *e.g.* [WAM02, YS06]) required a substantial amount of input, interaction, and/or expertise from a human operator who guides the algorithm (*e.g.* by providing scribbles of colour, choosing suitable reference images, segmenting images, or providing annotations). Later developments in this area sought to automate parts of the aforementioned user interactions with some success (see *e.g.* [Li08, Ch11]). Recently, fully automated methods based on massive datasets coupled with deep learning (*e.g.* [CYS15, ZIE16, He18]) and adversarial learning (*e.g.* [Ca17, NNE18]) emerged to address the limitations of previous methods specifically for the image colourisation task, and more generally for image-to-image translation problems (*e.g.* [Is17, Zh17]).

There exist different repositories with deep learning-based greyscale image colourisation software; however, many of them have certain limitations. For example, [E.17] only handles relatively low resolution images, while [Zh19] requires some user input in a semi-automatic process. In this work, we utilise one of the most recently published image colourisation methods called "DeOldify" [An19]. The software is based on concepts from [Zh18] and [He17], as well as a novel (as of yet unpublished) GAN pre-training strategy. In addition to the current version of the software, we also test an older version thereof [Ke19], which uses a different GAN training strategy inspired by [Ka17]. The authors of the software provide three pre-trained models: "artistic", "stable", and "video". We use the "stable" model, since according to the authors it is expected to achieve the best results *i.a.* for portraits, which are the use case considered in our paper. The used software which is applied to original images previously converted to greyscale[2] convinces with ex-

[2] Using ImageMagick command `magick in.png -grayscale Rec709Luma out.png`, see `https://imagemagick.org/script/command-line-options.php#grayscale`

cellent visual results and is easy and flexible to use. It should also be noted that this paper constitutes a preliminary study on this subject; future works (see section 5) may include considering a more comprehensive selection of image colourisation methods.

Figure 2 shows examples of colourised greyscale images (from the used facial image databases, see section 4.1) generated by the aforementioned methods. The results look mostly clean and realistic; furthermore, the newer model appears to produce more visually pleasing results. Note, that the colourised images are not identical to the corresponding original colour images. Those differences are inevitable: the colourisation algorithm needs to assign new pixel values in three dimensions (RGB) to pixel values with variation only along one dimension (intensity or luminance). It is possible for different colours to have the same luminance value, but different hue or saturation. Therefore, there exists no inherent "correct" solution to the task of colourising a greyscale image.

(a) original (b) greyscale (c) colourised [An19] (d) colourised [Ke19]

Fig. 2: Examples of reference (top row) and probe (bottom row) images: (a) the original images, (b) the original images converted to greyscale, (c)-(d) the colourised greyscale images.

3 Scenarios

As illustrated in figure 3, we consider five different scenarios which result in different pairings of compared reference and probe images:

- *Scenario 1*: baseline scenario; original reference and probe face images are used.
- *Scenario 2*: original reference image is used; probe image is converted to greyscale.
- *Scenario 3*: reference and probe images are converted to greyscale.
- *Scenario 4*: reference and probe images are converted to greyscale and colourised.
- *Scenario 5*: original reference image is used; probe image is converted to greyscale and colourised.

The second scenario might represent a surveillance or automated border control scenario in which a greyscale probe image is compared against a colour reference image. Accordingly, the third scenario reflects a use case where only greyscale images are processed by the face recognition system. Note that this applies for many older face recognition systems which utilise handcrafted feature extractors. The last two scenarios involve the application of colourisation to greyscale images. Specifically, colourisation is applied to reference and probe images or only to the probe image, respectively.

4 Experiments

The following subsections describe the experimental setup (section 4.1), conducted quality assessment (section 4.2), and biometric performance measurements (section 4.3).

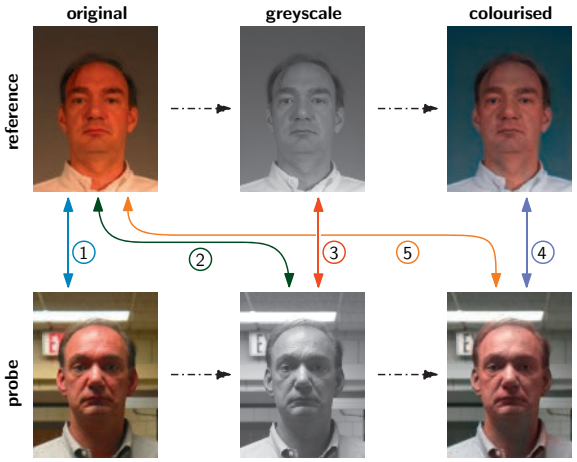


Fig. 3: Overview of applied the image processing chain (dash-dotted lines) and pairings (coloured lines) of reference and probe face images for the considered scenarios.

4.1 Experimental Setup

Subsets of two publicly available face image databases, *i.e.* FERET [Ph98] and FRGCv2 [Ph05], were used in the experiments. For reference images frontal faces with neutral expression have been manually chosen and ICAO compliance has been verified [In15]. Wherever possible, probe images from different acquisition session were preferentially chosen in order to obtain a realistic scenario. Examples of probe and reference images of both face image subsets are depicted in figure 2 and figure 3. The experiments are conducted in biometric verification mode. The number of subjects, corresponding reference and probe images, as well as the resulting number of genuine and impostor comparisons are listed in table 1.

Tab. 1: Overview of face image subsets from the FERET and FRGCv2 face databases.

Database	Subjects	Images		Comparisons	
		Reference	Probe	Genuine	Impostor
FERET	529	529	791	791	147,712
FRGCv2	533	984	1,726	3,298	144,032

For quality assessment, the FaceQnet algorithm [He19] is used.[3] This public face sample quality estimator is based on deep learning and returns a quality score (*i.e.* high values indicate good quality). In order to measure biometric performance, the widely used state-of-the-art ArcFace system [De19] is employed which has shown competitive recognition performance among open-source face recognition systems.[4] For a pair of reference and probe face images, this system returns a distance score (*i.e.* low values indicate high similarity). Note that when presented with a biometric sample, the face recognition system might internally perform some kind of colour space transformation(s). The utility of the individual colour channels for the purposes of face recognition has been investigated for

[3] FaceQnet has been shown to achieve convincing results, is open-source, and a pre-trained model is available, see `https://github.com/uam-biometrics/FaceQnet`

[4] ArcFace is open-source with a pre-trained model available at `https://github.com/deepinsight/insightface`

older systems by [BH08]. However, it is out of scope for this paper, as it investigates the effects of generative colourisation on facial recognition.

Tab. 2: Overview of face sample quality results.

Database	Colour	Mean	Median	St. Dev.	Minimum	Maximum
FERET	Original	0.616	0.616	0.050	0.460	0.777
	Greyscale	0.614	0.615	0.051	0.433	0.767
	Colourised [An19]	0.608	0.606	0.049	0.448	0.756
	Colourised [Ke19]	0.608	0.607	0.047	0.454	0.755
FRGCv2	Original	0.617	0.615	0.051	0.449	0.802
	Greyscale	0.622	0.620	0.053	0.426	0.794
	Colourised [An19]	0.616	0.614	0.053	0.435	0.777
	Colourised [Ke19]	0.615	0.613	0.052	0.453	0.811

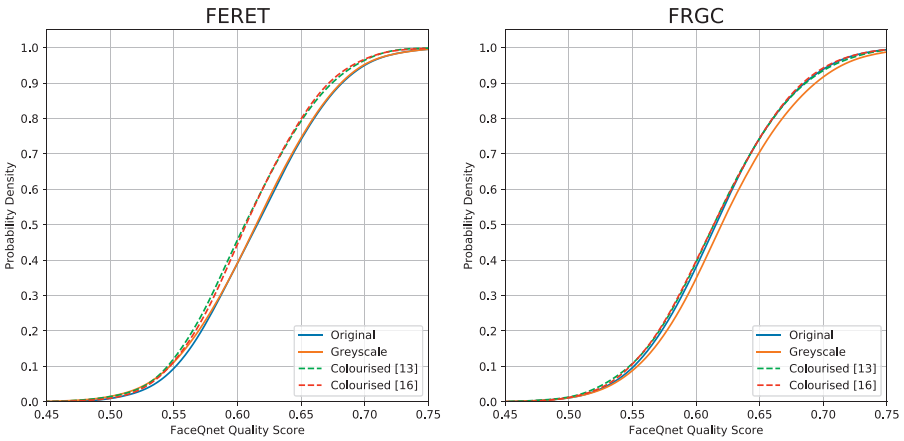
Biometric performance is evaluated in terms of False Non-Match Rate (FNMR) and False Match Rate (FMR). In addition, the Failure-to-Acquire rate (FTA) is measured as the proportion of verification attempts for which the system fails to capture or locate an image or signal of sufficient quality [IS06]. The False Reject Rate (FRR) is then estimated as the proportion of genuine verification transactions that will be incorrectly denied. This includes transactions denied due to failures-to-acquire as well as those denied due to false non-match decisions, $\mathrm{FRR} = \mathrm{FTA} + \mathrm{FNMR} \times (1 - \mathrm{FTA})$ [IS06]. More precisely, the FNMR and FRR are estimated at a false match probability of 0.1%, referred to as $\mathrm{FNMR}_{0.1}$ and $\mathrm{FRR}_{0.1}$, respectively. This operation point is recommended in the guidelines of European Agency for the Management of Operational Cooperation at the External Borders (FRONTEX) [FR15]. Genuine comparisons are performed for all of the previously described scenarios, while impostor comparisons are only performed for the first baseline scenario. That is, the decision thresholds estimated from the baseline scenario are used in all scenarios. Additionally, a decidability measure (d') [Da00] calculated as $d' = |\mu_g - \mu_i| / \sqrt{\frac{1}{2}(\sigma_g^2 + \sigma_i^2)}$ is reported, where μ_g and μ_i represent the means of the genuine and impostor score distributions and σ_g and σ_i their standard deviations, respectively.

4.2 Quality Assessment

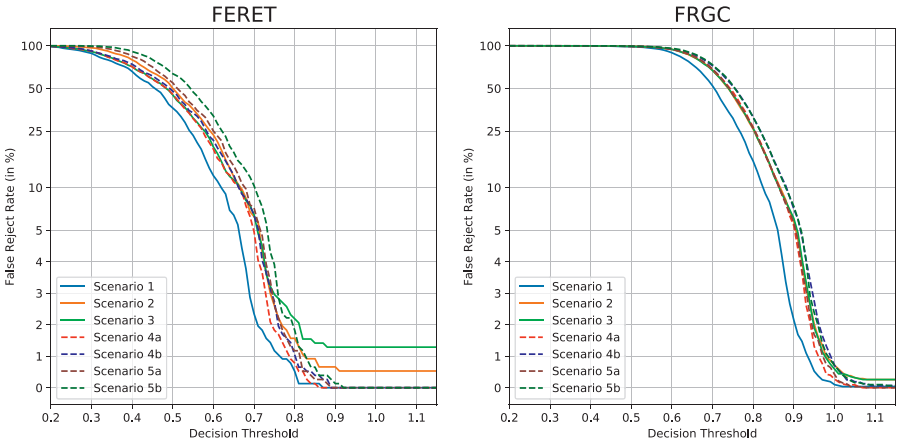
Figure 4(a) depicts the cumulative distribution function of sample quality scores. Corresponding statistical properties are summarised in table 2. The sample quality of the colourised images is generally lower compared to the greyscale images; however, the differences are very small. Hence, for the used method [He19], applied colourisation techniques do not yield improvements. This does not necessarily mean that colourisation might not be helpful in general – other algorithms in the pipeline might benefit from it, or it could be the case that other colourisation methods might improve the sample quality.

4.3 Performance Evaluation

Obtained biometric performance rates are summarised in table 3. On both databases practical biometric performance is achieved. Generally, better performance rates are obtained on the FERET database which contains more constrained face images. Further, it can be observed that increased FTAs are obtained for scenarios in which greyscale images are processed. However, by using image colourisation, the FTAs are reduced to that of the baseline system. That is, colourisation is found helpful to reduce the FTA on the greyscale



(a) Cumulative distribution of face sample quality.



(b) FRRs in relation to decision threshold.

Fig. 4: Experimental results.

images. This is also reflected in the FRRs which are plotted in figure 4(b) (the FRRs values begin at 100% for the low decision thresholds, as the used face recognition system works with dissimilarity-based comparison scores). Example face images for which the feature extraction failed on greyscale images but succeeded on the corresponding colourised images are shown in figure 5. Focusing on the algorithmic performance rates, *i.e.* FNMR, the application of colourisation yields generally worse comparison scores (higher dissimilarity) compared to scenarios in which greyscale images are processed directly.

5 Conclusion

Deep learning-based generative image colourisation techniques show impressive visual results for converting greyscale images to colour images. In this work, we investigated the usefulness of such techniques for facial recognition. For this purpose, open-source face image quality assessment and recognition tools are evaluated on two public databases

Tab. 3: Overview of biometric performance rates.

Database	Scenario	d'	FTA (%)	$FNMR_{0.1}$ (%)	$FRR_{0.1}$ (%)
FERET	1	9.290	0.000	0.000	0.000
	2	9.208	0.530	0.000	0.530
	3	8.793	1.288	0.000	1.288
	4 [An19]	8.739	0.000	0.000	0.000
	4 [Ke19]	8.553	0.000	0.000	0.000
	5 [An19]	9.344	0.000	0.000	0.000
	5 [Ke19]	9.306	0.000	0.000	0.000
FRGCv2	1	8.436	0.037	0.310	0.347
	2	7.872	0.258	0.446	0.703
	3	7.816	0.258	0.450	0.707
	4 [An19]	7.821	0.000	0.474	0.474
	4 [Ke19]	7.498	0.037	0.520	0.557
	5 [An19]	7.914	0.000	0.454	0.454
	5 [Ke19]	7.664	0.037	0.526	0.562

(a) greyscale (b) colourised [An19] (c) colourised [Ke19]

Fig. 5: Example face images for which the feature extraction failed on the greyscale images but succeeded on the colourised images.

considering scenarios where face images are converted to greyscale and colourised using state-of-the-art colourisation software.

In the conducted experiments, the effects of colourisation on sample quality were insignificant and did not result in improvements. To fully evaluate the impact of colourisation on biometric performance, more experiments with larger and more unconstrained datasets, as well as more facial recognition systems are needed to produce statistically significant results. The scenario-based evaluation of the comparison scores indicated generally inferior comparison scores for the colourised images compared to the direct use of greyscale images. These results are logically comprehensible since colourisation only aims at producing plausible colour images based on learned statistics which may vary for each image, *i.e.* colourised face images of various images a single subject may look different.

Finally, it was observed that colourisation can reduce the FTA, *i.e.* face detection and feature extraction exhibit more robustness if colourisation is applied. However, this may also highly depend on the used face recognition system. This preliminary study opens several avenues of potential research. Future works in this area may include testing other facial recognition and quality estimation methods, as well as different image colourisation schemes. Furthermore, the study could be extended to other applications of facial biometrics, such as biometric identification and classification of demographic attributes.

Acknowledgements

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

References

[An19] Antic, J.: , DeOldify – A Deep Learning based project for colorizing and restoring old images (and video!). `https://github.com/jantic/DeOldify`, 2019.

[BH08] Bours, P.; Helkala, K.: Face Recognition Using Separate Layers of the RGB Image. In: Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). IEEE, pp. 1035–1042, 2008.

[Ca17] Cao, Y.; Zhou, Z.; Zhang, W.; Yu, Y.: Unsupervised diverse colorization via generative adversarial networks. In: Machine Learning and Knowledge Discovery in Databases. Springer, pp. 151–166, 2017.

[Ca19] Cai, J.; Han, H.; Shan, S.; Chen, X.: FCSR-GAN: Joint Face Completion and Super-resolution via Multi-task Learning. Trans. on Biometrics, Behavior, and Identity Science (TBIOM), 2019.

[Ch11] Chia, A. Y.-S.; Zhuo, S.; Gupta, R. K.; Tai, Y.-W.; Cho, S.-Y.; Tan, P.; Lin, S.: Semantic colorization with internet images. In: Trans. on Graphics (TOG). volume 30. ACM, p. 156, 2011.

[Ch18] Chen, Y.; Tai, Y.; Liu, X.; Shen, C.; Yang, J.: FSRNet: End-to-End Learning Face Super-Resolution with Facial Priors. In: Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 2492–2501, 2018.

[CYS15] Cheng, Z.; Yang, Q.; Sheng, B.: Deep Colorization. In: Int. Conf. on Computer Vision (ICCV). IEEE, pp. 415–423, 2015.

[Da00] Daugman, J.: Biometric Decision Landscapes. Technical Report UCAM-CL-TR-482, University of Cambridge - Computer Laboratory, 2000.

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

[E.17] E.Wallner: , Coloring black and white images with deep learning. `https://blog.floydhub.com/colorizing-b-w-photos-with-neural-networks/`, 2017.

[FR15] FRONTEX: . Best Practice Technical Guidelines for Automated Border Control ABC Systems, 2015.

[GSŠ19] Grm, K.; Scheirer, W. J.; Štruc, V.: Face hallucination using cascaded super-resolution and identity priors. Trans. on Image Processing (TIP), pp. 2150–2165, 2019.

[Ha19] Ha, V. K.; Ren, J.-C.; Xu, X.-Y.; Zhao, S.; Xie, G.; Masero, V.; Hussain, A.: Deep Learning Based Single Image Super-resolution: A Survey. Int. J. of Automation and Computing (IJAC), 16(4):413–426, 2019.

[He17] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Int. Conf. on Neural Information Processing Systems (NIPS). ACM, pp. 6626–6637, 2017.

[He18] He, M.; Chen, D.; Liao, J.; Sander, P. V.; Yuan, L.: Deep exemplar-based colorization. Trans. on Graphics (TOG), 37(4):47, 2018.

[He19] Hernandez-Ortega, J.; Galbally, J.; Fiérrez, J.; Haraksim, R.; Beslay, L.: FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In: Int. Conf. on Biometrics (ICB). IEEE, 2019.

[In15] International Civil Aviation Organization: . Machine Readable Passports – Part 9 – Deployment of Biometric Identification and Electronic Storage of Data in eMRTDs, 2015.

[IS06] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework, 2006.

[Is17] Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A. A.: Image-to-image translation with conditional adversarial networks. In: Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1125–1134, 2017.

[Ka17] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[Ke19] Kelley, D.: , DeOldify – A Deep Learning based project for colorizing and restoring old images. `https://github.com/dana-kelley/DeOldify`, 2019.

[Li08] Liu, X.; Wan, L.; Qu, Y.; Wong, T.-T.; Lin, S.; Leung, C.-S.; Heng, P.-A.: Intrinsic colorization. In: Trans. on Graphics (TOG). volume 27. ACM, p. 152, 2008.

[Li17] Li, Y.; Liu, S.; Yang, J.; Yang, M.-H.: Generative Face Completion. In: Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3911–3919, 2017.

[Li18] Liu, G.; Reda, F. A.; Shih, K. J.; Wang, T.-C.; Tao, A.; Catanzaro, B.: Image Inpainting for Irregular Holes Using Partial Convolutions. In: European Conf. on Computer Vision (ECCV). Springer, pp. 85–100, 2018.

[Li19] Li, P.; Prieto, L.; Mery, D.; Flynn, P. J.: On Low-Resolution Face Recognition in the Wild: Comparisons and New Techniques. Trans. on Information Forensics and Security (TIFS), 14(8):2000–2012, 2019.

[LSF07] Liu, C.; Shum, H.-Y.; Freeman, W. T.: Face Hallucination: Theory and Practice. Int. J. of Computer Vision (IJCV), 75(1):115–134, 2007.

[MIA19] Mathai, J.; Iacopo, M.; Abd-Almageed, W.: Does Generative Face Completion Help Face Recognition? In: Int. Conf. on Biometrics (ICB). IEEE, 2019.

[NNE18] Nazeri, K.; Ng, E.; Ebrahimi, M.: Image colorization using generative adversarial networks. In: Int. Conf. on Articulated Motion and Deformable Objects. Springer, pp. 85–94, 2018.

[Ph98] Phillips, J.; Wechsler, H.; Huang, J.; Rauss, P.: The FERET Database and Evaluation Procedure for Face Recognition Algorithms. Image and Vision Computing Journal (IMAVIS), 16(5):295–306, 1998.

[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: Conf. on Computer Vision and Pattern Recognition (CVPR). volume 1. IEEE, pp. 947–954, 2005.

[WAM02] Welsh, T.; Ashikhmin, M.; Mueller, K.: Transferring color to greyscale images. In: Trans. on graphics (TOG). volume 21. ACM, pp. 277–280, 2002.

[YS06] Yatziv, L.; Sapiro, G.: Fast image and video colorization using chrominance blending. Trans. on Image Processing (TIP), 15(5):1120–1129, 2006.

[Zh17] Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Int. Conf. on computer vision (ICCV). IEEE, pp. 2223–2232, 2017.

[Zh18] Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318, 2018.

[Zh19] Zhang, R.; Zhu, J.-Y.; Isola, P.; Geng, X.; Lin, A. S.; Yu, T.; Efros, A. A.: , Interactive Deep Colorization. `https://github.com/junyanz/interactive-deep-colorization`, 2019.

[ZIE16] Zhang, R.; Isola, P.; Efros, A. A.: Colorful Image Colorization. In: European Conf. on Computer Vision (ECCV). Springer, pp. 649–666, 2016.

Gesellschaft für Informatik e.V. (GI)

publishes this series in order to make available to a broad public recent findings in informatics (i.e. computer science and information systems), to document conferences that are organized in cooperation 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 978-3-88579-700-5

The proceedings of the BIOSIG 2020 include scientific contributions of the annual international conference of the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik (GI). Due to the pandemic situation the conference was held as a digital conference, 16.-18. September 2020. 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.