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 978-3-88579-664-0

The proceedings of the BIOSIG 2017 include scientific contributions of the annual conference of the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik (GI). The conference took place in Darmstadt, 20.-22. September 2017. 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.

Arslan Brömme, Christoph Busch, Antitza Dantcheva, Christian Rathgeb, Andreas Uhl (Eds.):
BIOSIG 2017 - 16th International Conference of the Biometrics Special Interest Group

270

GI-Edition

Lecture Notes in Informatics

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

BIOSIG 2017

Proceedings of the 16th International Conference of the Biometrics Special Interest Group

20.–22. September 2017
Darmstadt, Germany

Proceedings

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

BIOSIG 2017

Proceedings of the 16th International Conference of the Biometrics Special Interest Group

**20.-22. September 2017 in
Darmstadt, Germany**

Gesellschaft für Informatik e.V. (GI)

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

Volume P-270
ISBN 978-3-88579-664-0
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 Méditerranée, 2004 Rue des
Lucioles, F-06902 Sophia Antipolis Cedex
Email: antitza.dantcheva@inria.fr

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. Over more than a decade 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 2017 international conference is jointly organized by the Biometrics Special Interest Group (BIOSIG) of the Gesellschaft für Informatik e.V., the Competence Center for Applied Security Technology e.V. (CAST), the German Federal Office for Information Security (BSI), the European Association for Biometrics (EAB), the European Commission Joint Research Centre (JRC), the TeleTrusT Deutschland e.V. (TeleTrusT), the Norwegian Biometrics Laboratory (NBL), the Center for Research in Security and Privacy (CRISP), Institution of Engineering and Technology Biometrics Journal Biometrics Journal (IET Biometrics), and the Fraunhofer Institute for Computer Graphics Research (IGD). This year's international conference BIOSIG 2017 is once again technically co-sponsored by the Institute of Electrical and Electronics Engineers (IEEE) and is enriched with satellite workshops by the TeleTrust Biometric Working Group and the European Association for Biometrics.

The international program committee accepted full scientific papers strongly according to the LNI guidelines (**acceptance rate ~33%**) within a scientific double-blinded review process of at minimum five reviews per paper. All papers were formally restricted for the printed proceedings up to 12 pages for regular research contributions including an oral presentation and up to 8 pages for further conference contributions including a poster presentation at the conference site.

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 2017 will be published additionally in the IEEE Xplore Digital Library.

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

Darmstadt, 20th September 2017

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

Chairs

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

Christoph Busch, *Hochschule Darmstadt, Germany*

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

Christian Rathgeb, *Hochschule Darmstadt, Germany*

Andreas Uhl, *University of Salzburg, Austria*

Program Committee

Harald Baier (HDA, DE)
Oliver Bausinger (BSI, DE)
Piotr Bilinski (Univ. Oxford, UK)
Thiriamchos Bourlai (WVU, US)
Patrick Bours (NTNU, NO)
Sebastien Brangoulo (Morpho, FR)
Andreas Braun (FHG IGD, DE)
Ralph Breithaupt (BSI, DE)
Julien Bringer (Morpho, FR)
Arslan Brömme (GI BIOSIG, DE)
Christoph Busch (CAST-Forum, DE)
Victor-Philipp Busch (Sybuca, DE)
Patrizio Campisi (Uni Roma 3, IT)
Cunjian Chen (MSU, US)
Nathan Clarke (CSCAN, UK)
Adam Czajka (NASK,PL)
Farzin Deravi (UKE, UK)
Antitza Dantcheva (INRIA, FR)
Martin Drahansky (BUT, CZ)
Andrzej Drygajlo (EPFL, CH)
Julian Fierrez (UAM, ES)
Lothar Fritsch (KAU, SE)
Steven Furnell (CSCAN, UK)
Sonia Garcia (TSP, FR)
Marta Gomez-Barrero (HDA, DE)
Patrick Grother (NIST, US)
Olaf Henniger (FHG IGD, DE)
Heinz Hofbauer (COSY, AT)
Stan Li (CBSR, CN)
Paulo Lobato Correira (IST, PT)
Davide Maltoni (UBO, IT)
Johannes Merkle (secunet, DE)
Emilio Mordini (CSSC, IT)
Kamal Nasrollahi (AAU, DK)
Mark Nixon (UoS, UK)
Alexander Nouak (Fraunhofer, DE)
Markus Nuppeney (BSI, DE)
Hisao Ogata (Hitachi, JP)
Martin Olsen (Fingerprints, NO)
Michael Peirce (Daon, IR)
Dijana Petrovska (TSP, FR)
Kiran Raja (NTNU, NO)
Raghu Ramachandra (NTNU, NO)
Kai Rannenberg (Uni FFM, DE)
Christian Rathgeb (HDA, DE)
Arun Ross (MSU, US)
Heiko Roßnagel (FHG IAO, DE)
Raul Sanchez-Reillo (UC3M, ES)
Stephanie Schuckers (ClU, US)
Günter Schumacher (JRC, IT)
Takashi Shinzaki (Fujitsu, JP)
Luis Soares (ISCTE-IUL, PT)
Luuk Spreeuwers (UTW, NL)
Syed Zulkarnain Syed Idrus (UniMAP, MY)
Tieniu Tan (NLPR, CN)
Massimo Tistarelli (UNISS, IT)

Abdenour Hadid (UO, FI)
Detlef Hühnlein (ecsec, DE)
Stefan Katzenbeisser (TUD, DE)
Tom Kevenaar (GenKey, NL)
Ulrike Korte (BSI, DE)
Ajay Kumar (Poly, HK)
Young-Bin Kwon (CAU, KR)
Herbert Leitold (a-sit, AT)
Guoqiang Li (NTNU, NO)
Dimitrios Tzovaras (CfRaT, GR)
Andreas Uhl (COSY, AT)
Markus Ullmann (BSI, DE)
Raymond Veldhuis (UTW, NL)
Jim Wayman (SJSU, US)
Peter Wild (AIT, AT)
Andreas Wolf (BDR, DE)
Bian Yang (NTNU, NO)

Hosts

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

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

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

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

European Commission Joint Research Centre (**JRC**)
http://ec.europa.eu/dgs/jrc/index.cfm

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

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

Center for Research in Security and Privacy (**CRISP**)
https://www.crisp-da.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 2017 – Biometrics Special Interest Group

"**2017 International Conference of the Biometrics Special Interest Group**"
20th -22nd September 2017

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 2017 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 2017

Regular Research Papers

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 15

Resting-state EEG: A Study on its non-Stationarity for Biometric Applications

Gabriel Emile Hine, Emanuele Maiorana, Patrizio Campisi [1]

Abstract: In the last years, several papers on EEG-based biometric recognition systems have been published. Specifically, most of the proposed contributions focus on brain signals recorded in resting state conditions, with either closed or open eyes. A common assumption is that the acquired signals are quasi-stationarity. In this paper, we investigate such property in terms of discriminative capability, and we analyze whether or not it holds throughout the entire duration of data collected over long periods. An extensive set of experimental tests, conducted over a database comprising signals collected from 50 subjects in three distinct acquisition sessions, shows that the most distinctive information of the brain signals is temporally located at the beginning of each recording.

Keywords: EEG, Biometrics, non-Stationarity.

1 Introduction

Brain signals, after having been investigated in the medical field since the beginning of the twentieth century, have recently attracted the attention of the scientific community as biometric identifiers, to be used in automatic people recognition systems [PM07]. In fact, it has been postulated that our brain possesses peculiar subject-specific properties, whose analysis could efficiently allow discriminating between distinct persons [Ri08]. Within this framework, among the different modalities that can be used to sense brain's activity, electroencephalography (EEG) has received most of researchers' interest, since it permits collecting brain information using portable and relatively inexpensive devices, a notable advantage to foster the adoption of such trait in practical biometric recognition systems [Kl13].

EEG signals are sensed through electrodes placed on the head scalp surface as voltage differences of the electrical field generated by the synchronous firing of specific spatially-aligned neurons of the cortex, i.e., the pyramidal neurons [Ga77]. The characteristics of the collected data commonly depend on the acquisition protocol employed to elicit a specific subjects' behavior. Different task-related responses, typically represented in terms of small time-locked changes in the electrical activity of the brain, can in fact be obtained using stimulation paradigms involving various sensory, cognitive or motor stimuli [CLR14].

However, the elicitation protocol most commonly employed in both medical- and biometrics–related studies simply requires the considered subjects to remain in a relaxed, yet vigilant, state, in either eyes-closed (EC) or eyes-open (EO) conditions. Since EEG signals thus acquired do not contain characteristics related to some specific event but the

[1] Section of Applied Electronics, Department of Engineering, Roma Tre University, Rome, Italy,
{gabriel.hine, emanuele.maiorana, patrizio.campisi}@uniroma3.it

beginning of the recording, they are typically processed by splitting the available data into multiple (possibly overlapping) epochs of a given duration [NWS07], from which representative features can be then extracted. The common assumption underlying this approach is that resting-state EEG data exhibit a quasi-stationary behavior, at least for the considered epoch length [Bl95]. Nevertheless, it is also widely agreed that brain signals are inherently non-stationary over long time periods [Al14], due to the properties of the underlying neural processes [Jo12]. Within medical literature, several statistical studies have investigated the extent of the time interval where the EEG stationarity assumption may hold [CG10], with controversial results estimating amounts ranging from several seconds to several minutes [Ra16].

On the other hand, to the best of our knowledge, no work has so far explicitly assessed the effects of brain signals non-stationarity on the distinctive characteristics of EEG data used for biometric applications. Such aspect has in fact been only partially taken into account in [MLRC16], where the permanence of EEG discriminative traits over a period of one month, implicitly connected to long-term stationarity of brain signals, has been analyzed.

Actually, it would be extremely important to understand if variable discriminative properties are exhibited by distinct epochs extracted during an EEG recording, that is, if EEG non-stationarity affects the characteristics of epochs taken at different time distances from the beginning of a resting-state acquisition session. This issue, besides providing interesting insights on the properties of the neural processes generating EEG signals, could guide the optimization of both the enrolment and the recognition stages in EEG-based biometric recognition systems, from an applicative point. In fact, determining that EEG non-stationarity affects the achievable recognition performance could for instance suggest to limit the length of both enrolment and recognition phases to intervals where the available data could be considered stationary, in order to allow the extraction of coherent discriminative features. Moreover, depending on the results of such analysis, it may also be advisable to perform continuous recognition schemes using resting-state EEG biometrics only in case the available recordings could be separated into multiple instances, through the periodic presentation of specific stimuli to the considered subjects.

The proposed paper is organized as follows: the biometric recognition system considered to evaluate the effects of EEG non-stationarity on the achievable accuracy performance is presented in Section 2, while the employed EEG database is outlined in Section 3. The performed experimental tests, together with the obtained results, are then discussed in Section 4, with the deriving conclusions drawn in Section 5.

2 Employed Recognition System

The present section describes the biometric system employed to evaluate the effects of EEG non-stationarity on the achievable recognition accuracy. The considered system is designed according to the best performing architecture evaluated in [MLRC16]. In more detail, let us assume that an M-channel EEG system is available for either enrolment or identification purposes. The preprocessing applied to the available EEG data comprises:

- common average referencing (CAR) spatial filtering, applied to reduce artifacts related to unsuitable reference choices or not-expected reference variations [Mc97];
- band-pass filtering in the EEG range $\{\alpha, \beta\}$ = [8,30] Hz, carrying the most discriminative information according to the performed experimental tests;
- down-sampling from the original rate to 64 Hz, in order to reduce the computational complexity of the subsequent processing;
- segmentation in epochs lasting $L = 5$ s, with an overlapping factor of O = 40%. It is implicitly assumed that the treated EEG signals show a stationary behavior along the considered epoch length.

After preprocessing, discriminative features are extracted from each signal available in a given EEG epoch. The considered features are the Burg's reflection coefficients of an autoregressive (AR) model, which have proven to allow achieving the best possible recognition performance when exploited for representing EEG data in [MLRC16]. In more detail, an AR model of order Q =12 is here employed for template generation, to minimize the information loss in fitting the considered data according to the Akaike information criterion (AIC). The reflection coefficients extracted from each EEG channel are then concatenated to form a single feature vector $\mathbf{v}$ with length $B = Q \cdot M$ to represent the whole epoch.

Having indicated with $\mathbf{v}_e^u[b]$, $b = 1,\ldots,B$ and $e = 1,\ldots,E$, the representation obtained from the e-th epoch of user u's enrolled EEG data, and with $\mathbf{v}_i[b]$, $i = 1,\ldots,I$ the template associated to the i-th epoch of the probe EEG signal submitted during an identification attempt, matching the two considered EEG recordings requires computing the I scores

$$\begin{aligned} D_i^u &= \min_e d_{i,e}^u, \\ d_{i,e}^u &= \sum_{b=1}^{B} |\mathbf{v}_i[b] - \mathbf{v}_e^u[b]|, \; i = 1,\ldots,I, \end{aligned} \tag{1}$$

through a Manhattan (L1) distance metrics. Such values can be then used to estimate a potential identity $\hat{u}_i = \arg\min_u \{D_i^u\}$ for each available i-th identification epoch. The final decision $\hat{u}$ is taken according to a majority voting rule, selecting the identity with the highest number of occurrences among the votes $\hat{u}_i$, $i = 1,\ldots,I$.

The tests described in Section 4 leverage both on the distributions of genuine scores D_i^u, as well as on the overall identification rate achievable with the outlined recognition system, to assess the effects of signals non-stationarity on EEG discriminative capabilities.

3 EEG Database

The EEG database exploited for testing has been collected in the authors' lab. The database contains data taken from 50 healthy subjects, whose age ranges from 20 to 35 years with an average of 25. EEG signals are acquired at an original sampling rate of 256 Hz through $M = 19$ electrodes placed on the scalp according to the 10-20 international system [MP95],

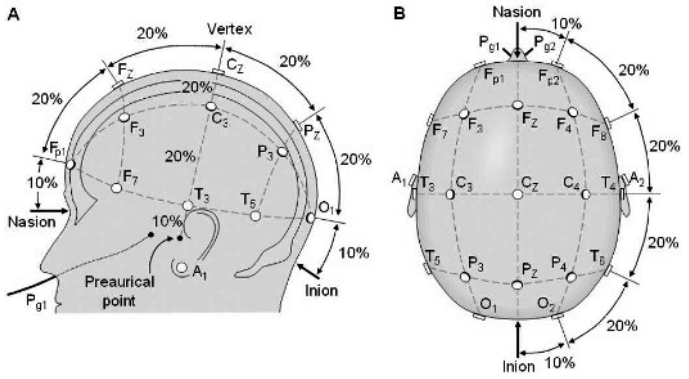


Fig. 1: The 10-20 International system seen from left (A) and above the head (B), from Jaakko Malmivuo and Robert Plonsey, Bioelectromagnetism, Oxford University Press, 1995, WEB version).

as shown in Figure 1. Three different EEG recordings, indicated in the following as S1, S2 and S3, acquired during three distinct sessions spanning a period of approximately one month, are available for each subject. During each session, EEG signals are first recorded for four minutes in EO conditions, while fixing a small spot of light on a screen. EEG signals are then taken for other additional 4 minutes in an EC scenario.

The experimental tests described in Section 4 are performed by comparing EEG data acquired in different sessions. It has to be remarked that, although rarely followed in literature, such approach is actually the preferable one to be used for properly evaluating the discriminating characteristics of brain signals [MLRC16]. The vast majority of studies dealing with EEG biometrics in fact performs tests on data collected during a single acquisition session, leveraging on different partitions of the available data to generate training and testing samples [CLR14]. Nevertheless, the reliability of such methodology is questionable, since it is hard to state whether the reported recognition performance depends only on the characteristics of each subject's neural activity, or also on session-specific exogenous conditions, such as the capacitive coupling of electrodes and cables with lights or computer, induction loops between the employed equipment and the body, power supply artifacts, and so on. These latter may in fact significantly differ between distinct acquisition sessions, thus affecting both inter- and intra-class variability of EEG recordings.

4 Experimental Results

The first experimental test conducted to verify the influence of EEG non-stationarity on brain signals discriminative capabilities is performed taking into account the scores $d^{u}_{i,e}$ obtained comparing the e-th epoch made available during each subject u's enrolment, and the i-th epoch extracted from the identification probe. Six different scenarios are considered to exploit the EEG data recorded from the $U = 50$ available subjects, using signals captured during session S1 for enrolment and those recorded in session S2 as identification data (S1 *vs* S2), and analogously S1 *vs* S3, S2 *vs* S1, S2 *vs* S3, S3 *vs* S1, and S3 *vs* S2. For each

$\{i,e\}$ correspondence between enrolment and identification epochs, 50x6=300 genuine scores and 50x49x6 = 14700 impostor scores are evaluated on the basis of the distances $d^u_{i,e}$, $u = 1,\ldots,U$ computed over the six considered scenarios. Distributions of genuine scores are then characterized through their mean $\mu^{(G)}_{i,e}$ and standard deviation $\sigma^{(G)}_{i,e}$, as well as distributions of impostor scores, represented with the associated mean $\mu^{(F)}_{i,e}$ and standard deviation $\sigma^{(F)}_{i,e}$. The distinctiveness of EEG signals is then evaluated through d-prime measure [BPR00]:

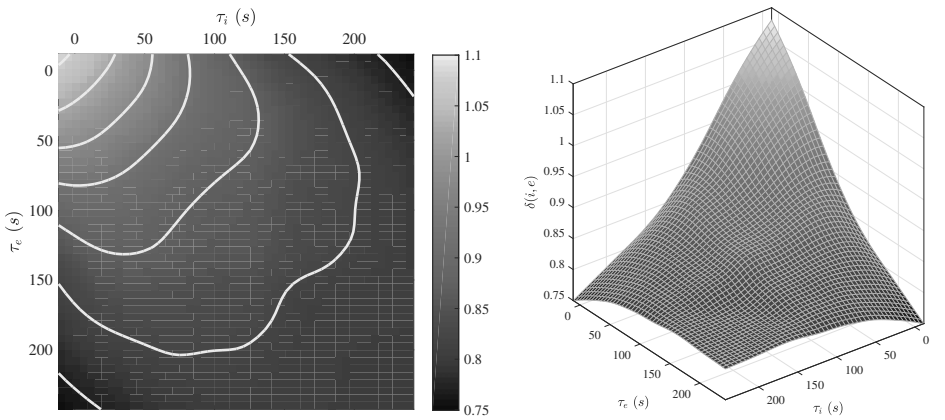
$$\delta(i,e) = \frac{|\mu^{(G)}_{i,e} - \mu^{(F)}_{i,e}|}{\sqrt{(\sigma^{(G)}_{i,e})^2 + (\sigma^{(F)}_{i,e})^2}}, \tag{2}$$

for each considered $\{i,e\}$ couple of enrolment and identification epochs. Larger values of this metric indicate higher discriminative capabilities, being obtained from larger differences between the means of genuine and impostor scores, or smaller variances of the computed distributions. Figures 2 and 3 show the behavior of $\delta(i,e)$ in EC and EO conditions, with respect to the time distances from the beginning of enrolment and identification sessions τ_e and τ_i at which the compared epochs are selected, with $\tau_e = L\cdot[1+(e-1)\cdot(1-O)]$ and $\tau_i = L\cdot[1+(i-1)\cdot(1-O)]$. The locally weighted scatter plot smoothing (LOWESS) method [Cl79] is applied to the obtained data in order to show a well-defined trend, for both the aforementioned and following experimental tests.

From the shown figures it is possible to observe that, in both the considered resting state protocols, processing epochs close to the beginning of each acquisition allows to exploit EEG characteristics much more discriminative than those available at a later time. The amount of subject-specific information present in a single epoch of EEG signals acquired in resting conditions therefore seems to diminish as long as the length of recording sessions increases.

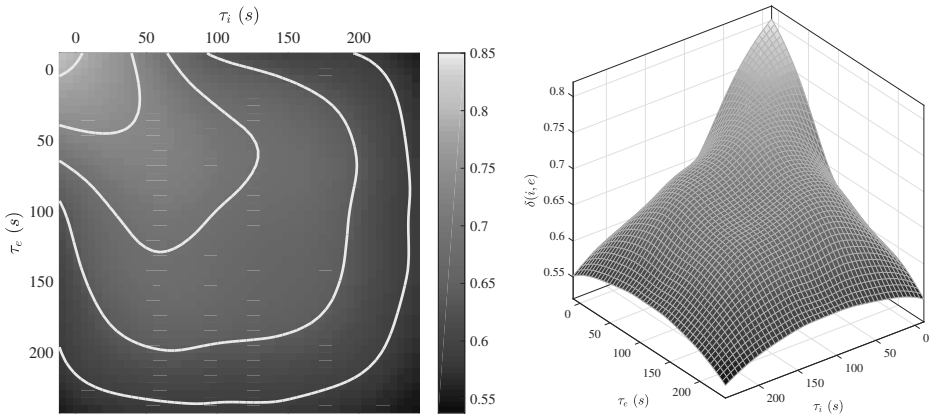
The effects of such behavior on the rank-1 identification performance achievable when employing the recognition system described in Section 2 are outlined in Figures 4 and 5 for the two considered resting state protocols. The obtained correct recognition rates (CRR) show rapid improvements when enrolment and identification phases initially increase in length, yet only a negligible improvement is obtained when the employed phases last more than a couple of minutes, approximately.

Further evidence of the presence of more discriminative characteristics in epochs close to the beginning of an EEG acquisition is obtained when evaluating the identification rates achievable comparing signals of fixed length, yet taken at different starting offsets from the actual beginning of a resting-state EEG recording session. Figures 6 and 7 show the CRRs obtained in this experiment, when using enrolment and identification signals lasting 90 *s*. Also this test outlines that better identification rates can be attained when considering EEG signals starting at offsets close to the actual beginning of the performed recording sessions. In more detail, we can appreciate that the identification offset affects the achievable performances much more than the enrolment offset does. Such behavior is however specifically due to the employed identification scheme, which searches in (1) for the min-



(a) (b)

Fig. 2: Distinctiveness of EEG in EC resting states, expressed in terms of the measure δ, evaluated comparing epochs extracted at different time distances τ_e and τ_i from the beginning of enrolment and identification recording sessions. (a): surface plot with contour lines; (b) mesh plot.



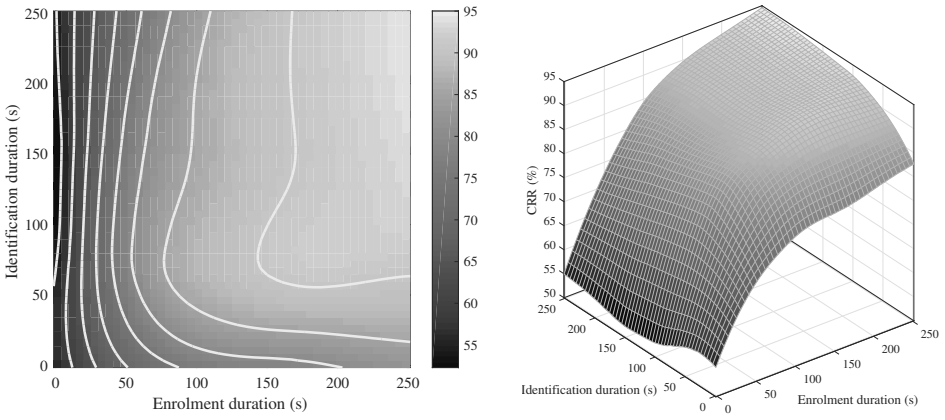
(a) (b)

Fig. 3: Distinctiveness of EEG in EO resting states, expressed in terms of the measure δ, evaluated comparing of epochs extracted at different time distances τ_e and τ_i from the beginning of enrolment and identification recording sessions. (a): surface plot with contour lines; (b) mesh plot.

imum L1 distance from the specific identification probe and each of the epochs available in the enrolment set.

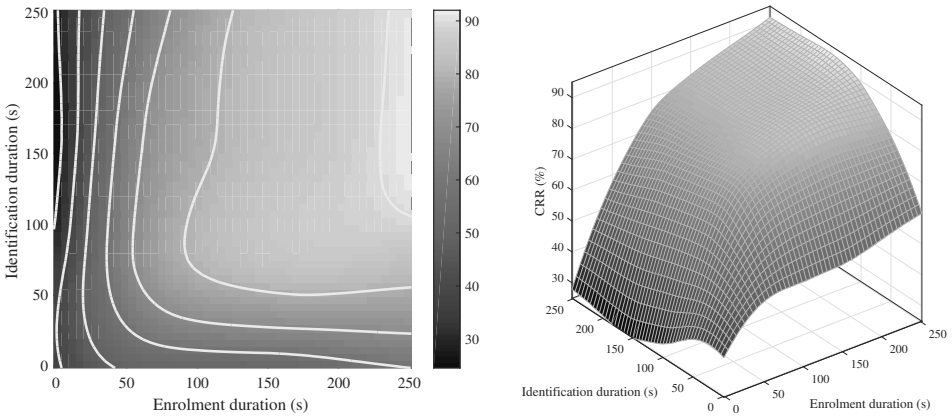
5 Conclusions

The present paper has investigated the effects of non-stationarity for EEG signals collected according to resting state protocols with either EC or EO conditions on their discriminative capabilities, when such data are exploited as biometric identifiers. The reported exper-



(a) (b)

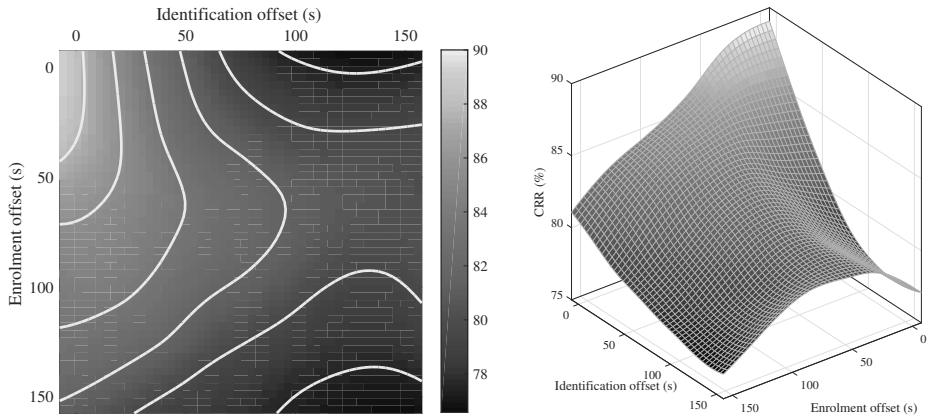
Fig. 4: CRR vs enrolment and identification durations, in EC resting states. (a): surface plot with contour lines; (b) mesh plot.



(a) (b)

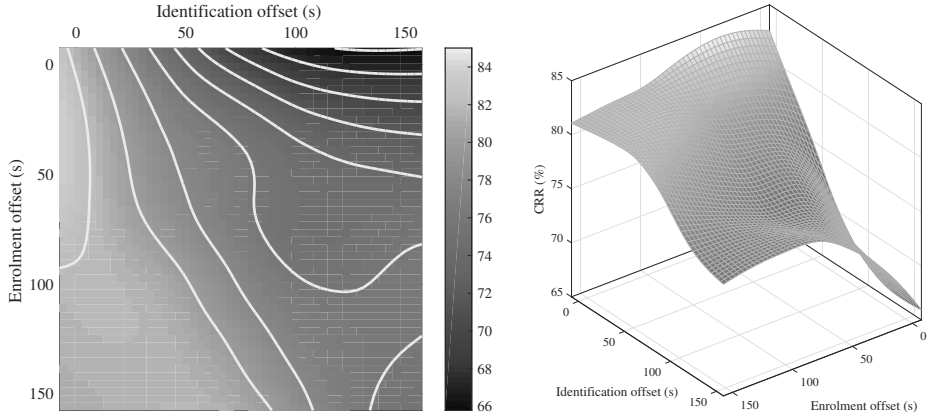
Fig. 5: CRR vs enrolment and identification durations, in EO resting states. (a): surface plot with contour lines; (b) mesh plot.

imental tests, executed on a database comprising recordings taken from 50 users during three distinct sessions, highlight that the initial part of an EEG acquisition performed in resting state conditions contains most of the discriminative characteristics offered by the considered biometrics. As an EEG recording is carried on, the acquired signals hold less subject-specific information, being their relevance for biometric purposes diminished. The obtained results suggest to not perform EEG recordings lasting more than a couple of minutes for either enrolment or identification purposes. In case a continuous recognition framework should be realized, it would be required to include some intermission during the procedure, in order to divide the acquired EEG signal into multiple instances of reduced duration, beginning from a repeated starting stimulus.



(a) (b)

Fig. 6: CRR vs enrolment and identification offsets (90-*s* duration), in EC resting states. (a): surface plot with contour lines; (b) mesh plot.



(a) (b)

Fig. 7: CRR vs enrolment and identification offsets (90-*s* duration), in EO resting states. (a): surface plot with contour lines; (b) mesh plot.

References

[Al14] Allen, Elena A.; Damaraju, Eswar; Plis, Sergey M.; Erhardt, Erik B.; Eichele, Tom; Calhoun, Vince D.: Tracking Whole-Brain Connectivity Dynamics in the Resting State. Cerebral Cortex, 24(3):663, 2014.

[Bl95] Blanco, S.; Garcia, H.; Quiroga, R. Q.; Romanelli, L.; Rosso, O. A.: Stationarity of the EEG series. IEEE Engineering in Medicine and Biology Magazine, 14(4):395–399, Jul 1995.

[BPR00] Bolle, R. M.; Pankanti, S.; Ratha, N. K.: Evaluation techniques for biometrics-based authentication systems (FRR). In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. volume 2, pp. 831–837 vol.2, 2000.

[CG10] Chang, Catie; Glover, Gary H: Time–frequency dynamics of resting-state brain connectivity measured with fMRI. Neuroimage, 50(1):81–98, 2010.

[Cl79] Cleveland, William S: Robust locally weighted regression and smoothing scatterplots. Journal of the American statistical association, 74(368):829–836, 1979.

[CLR14] Campisi, P.; La Rocca, D.: Brain waves for automatic biometric-based user recognition. IEEE Transactions on Information Forensics and Security, 9(5):782–800, 2014.

[Ga77] Gasser, T.: General characteristics of the EEG as a signal. In (Remond, A., ed.): EEG informatics: a Didactic Review of Methods and Applications of EEG Data Processing, pp. 37–55. Elsevier, 1977.

[Jo12] Jones, David T; Vemuri, Prashanthi; Murphy, Matthew C; Gunter, Jeffrey L; Senjem, Matthew L; Machulda, Mary M; Przybelski, Scott A; Gregg, Brian E; Kantarci, Kejal; Knopman, David S et al.: Non-stationarity in the resting brain's modular architecture. PloS one, 7(6):e39731, 2012.

[Kl13] Klonovs, J.; Petersen, C.K.; Olesen, H.; Hammershoj, A.: ID Proof on the Go: Development of a Mobile EEG-Based Biometric Authentication System. IEEE Vehicular Technology Magazine, 8(1):81–89, 2013.

[Mc97] McFarland, D.; McCane, L.; David, S.; Wolpaw, J.: Spatial filter selection for EEG-based communication. Electroencephalography and Clinical Neurophysiology, 103(3):386–394, September 1997.

[MLRC16] Maiorana, Emanuele; La Rocca, Daria; Campisi, Patrizio: On the Permanence of EEG Signals for Biometric Recognition. IEEE Transactions on Information Forensics and Security, 11(1):163–175, 2016.

[MP95] Malmivuo, J.; Plonsey, R.: Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields. Oxford University Press, 1995.

[NWS07] Napflin, M.; Wildi, M.; Sarnthein, J.: Test-retest reliability of resting EEG spectra validates a statistical signature of persons. Clinical Neurophysiology, 118(11):2519 – 2524, 2007.

[PM07] Palaniappan, R.; Mandic, D.P.: Biometrics from Brain Electrical Activity: A Machine Learning Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(4):738–742, 2007.

[Ra16] Rasoulzadeh, V.; Erkus, E. C.; Yogurt, T. A.; Ulusoy, I.; Zergeroğlu, S. Aykan: A comparative stationarity analysis of EEG signals. Annals of Operations Research, pp. 1–25, 2016.

[Ri08] Riera, A.; Soria-Frisch, A.; Caparrini, M.; Grau, C.; Ruffini, G.: Unobtrusive biometric system based on electroencephalogram analysis. EURASIP Journal of Advances in Signal Processing, 2008(143728), 2008.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 25

Multimodal Neural Network for Overhead Person Re-identification

Aske R. Lejbølle[1], Kamal Nasrollahi[2], Benjamin Krogh[3], Thomas B. Moeslund[4]

Abstract: Person re-identification is a topic which has potential to be used for applications within forensics, flow analysis and queue monitoring. It is the process of matching persons across two or more camera views, most often by extracting colour and texture based hand-crafted features, to identify similar persons. Because of challenges regarding changes in lighting between views, occlusion or even privacy issues, more focus has turned to overhead and depth based camera solutions. Therefore, we have developed a system, based on a Convolutional Neural Network (CNN) which is trained using both depth and RGB modalities to provide a fused feature. By training on a locally collected dataset, we achieve a rank-1 accuracy of 74.69%, increased by 16.00% compared to using a single modality. Furthermore, tests on two similar publicly available benchmark datasets of TVPR and DPI-T show accuracies of 77.66% and 90.36%, respectively, outperforming state-of-the-art results by 3.60% and 5.20%, respectively.

Keywords: Multimodal; Person Re-identification; Convolutional Neural Networks; Feature Fusion

1 Introduction

Person re-identification (re-id) i.e. identifications of persons across two or more cameras, is a topic with increasing interest due to potential usage in forensics, analysis of pedestrian flow in urban areas or monitoring of queue times in, for example, an airport. Meanwhile, it is also a topic still in research due to challenges that include changes in lighting, view and pose between camera views. To cope with these challenges, focus often lies in extracting robust hand-crafted feature descriptors from each person that are matched between views. For this purpose, soft biometrics are considered, such as colour and texture of the clothing, either represented as histograms [Li15] or transformed to sparse descriptors [LSF15]. To further improve accuracy of correct matches, supervised learning algorithms are applied that learn to separate similar feature pairs from dissimilar ones [Ch16, ZXG16]. More recently, deep learning has drawn increasing interest from the research community with Convolution Neural Networks (CNN) outperforming hand-crafted feature descriptors, as they are able to learn more expressive features [AJM15, WCZ16].

Besides aforementioned challenges, privacy preservation is often related to person re-id as a potentially large amount of data needs to be stored. Other than representing images as

[1] Visual Analysis of People (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark, asrl@create.aau.dk
[2] Visual Analysis of People (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark, kn@create.aau.dk
[3] BLIP Systems A/S, Vester Hassing, Haekken 2, 9310 Vodskov, Denmark, bbk@blipsystems.com
[4] Visual Analysis of People (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark, tbm@create.aau.dk

feature descriptors, camera placement can be considered as a means of privacy preservation. Most current benchmark datasets within re-id consider a frontal view [GT08, Li14] while only few consider an overhead view which has the advantage of reducing privacy issues and avoid occlusions between persons or objects and persons in the scene [HAFF16, Li17]. Furthermore, other modalities that are more anonymous can be used, for example depth, from which information is captured using either passive stereo, i.e. a stereo camera or active, for example, a Microsoft Kinect. From depth information, the height and width of the person can be extracted along with different body ratios [Ba12]. Instead of relying on a single modality, combining (fusing) different modalities have shown to improve performance in related applications such as object recognition [Ei15] and object segmentation [Ha16]. Such fusing can be done either at feature level (feature fusion), for example, by concatenation of respective feature descriptors or at decision/score level (late fusion) by fusing the output decisions/scores from different modalities [Ki98].

To consider challenges regarding changes across views and the advantages of fusing different modalities, we propose a novel framework for applying colour and depth (RGB-D) based re-id to images, captured with an overhead view. More specifically, we take advantage of the recent advances within deep learning and train a CNN using information from both RGB and depth modalities to improve accuracy compared to using either modality independently. To that extend, we collect a novel RGB-D based dataset in an uncontrolled environment from a stereo camera placed overhead to avoid occlusions and, at the same time, preserve privacy by not recording faces. Our dataset is collected to resemble real-life situations by having multiple persons within view, while current overhead datasets only consider a single person within view at a time. In summary, the main contributions of our work include:

- We train a CNN using RGB and depth modalities information and show that fusion of these improves accuracy.
- We collect and annotate a novel RGB-D and overhead based dataset which can be used to both evaluate re-id accuracy but also multi-target detection and tracking algorithms in RGB and depth domain.

2 Related Work

While re-id using hand-crafted colour and texture features or CNN's are widely studied, overhead re-id is rarely considered. In addition, only a limited number of articles suggest depth modality for this purpose.

Overhead re-id As most current re-id datasets are collected in outdoor scenes, a frontal view is typically considered. A few systems have been proposed for evaluating datasets with an overhead view [Ar08, AC12]. [AC12] proposes feature extraction using a Histogram of Oriented Gradients (HOG) algorithm combined with a linear Support Vector Machine (SVM) for classification while [Ar08] extracts features based on the colour and texture of the hair. While both datasets are recorded in an indoor environment, they only extract colour information.

Overhead RGB-D-based re-id More RGB-D based datasets for re-id are currently being proposed. While the first considered a frontal view [Mu14], the most recent consider an overhead view [HAFF16, Li17]. [HAFF16] collected a dataset in a hallway and applies a combined CNN and Long-Short-Term-Memory (LSTM) network using depth based image sequences to learn spatio-temporal representations of each person. Meanwhile, [Li17] extracts seven different depth features and two colour features that are feature fused by concatenation. While the former extracts only depth information, the latter considers only hand-crafted features from both modalities.
Multi-modal CNN While the work of [HAFF16] to our knowledge is the only previous proposed neural network using depth information for re-id, multi-modal CNN's have been proposed for related applications [Ei15, Ha16]. [Ei15] trains a CNN for object recognition using both colour and depth images by fusing respective features in late layers of the network to consider both modalities during training. To that extend, [Sa16] shows that feature fusion of colour and depth features in a CNN outperforms similar fusion scheme using other classification methods, such as SVM and Deep Belief Networks (DBN). Meanwhile, [Ha16] proposes a multi-modal encoder-decoder network for semantic segmentation by fusing outputs from each layer in an RGB and depth based encoder, respectively, before passing the output through an RGB-D based decoder. In this case, fusing is applied as an element-wise summation. To our knowledge, no multi-modal neural networks have previously been proposed for re-id. Although, [WCZ16] proposes a fusing scheme similar to that of [Ei15], but instead of fusing different modalities, complementary feature types are fused, i.e. CNN and hand-crafted features. To our knowledge, the system proposed in this paper, is the first to incorporate multiple modalities in a CNN to learn a multi-modal feature representation.

3 Methodology

As we desire to exploit both colour and depth information, along with the potential of CNN's, our aim is to use an architecture which jointly processes the two modalities, RGB and depth, simultaneously. For person re-id, such architecture has not previously been applied, although, in object recognition the work of [Ei15] shows an increase in accuracy compared to using a single modality.

We apply an architecture similar to that of [Ei15], having two CNN streams separately processing an input image while being fused in a later fully connected layer, as shown in Figure 1. The structure of each separate CNN follows the AlexNet architecture (please see [KSH12] for details) and consists of five convolution layers, the first, second and fifth followed by a max pooling and normalization layer. The outputs from the last convolution layers are followed by two fully connected layers, transforming the feature maps to sparse representations for RGB and depth, respectively. The feature representations ($fc7^{RGB}$ and $fc7^{D}$) are concatenated and used as input to a fully connected layer (fc8) which learns a joint RGB-D feature representations based on both colour and depth images. Finally, a softmax layer (fc9) calculates output probabilities for each class, defined as a person ID, which combined with a loss function is used to update the parameters of the network. We

refer to our proposed system as RGB-D-CNN. At test time, the softmax layer is discarded and features are extracted from fc8.

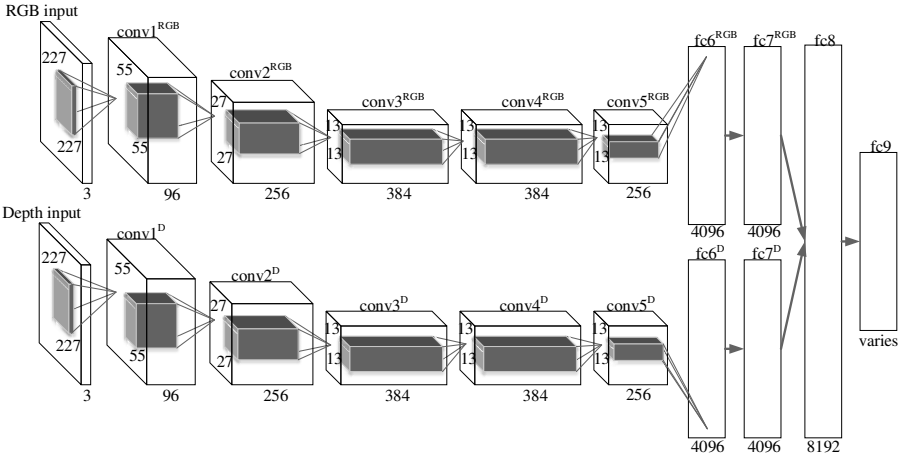


Fig. 1: Overview of the RGB-D based CNN (RGB-D-CNN). Lower part processes a depth image, while the upper part processes a colour image, features from last fully connected layer of the streams are fused in a joint fully connected layer before classification.

Individual training Before training the RGB-D-CNN, CNN models are trained for RGB and depth, respectively. We refer to these models as RGB-CNN and D-CNN. Both follow similar structure as the upper/lower part of the joint CNN, with a softmax layer replacing fc8 and fc9. The model weights are initialized using a pre-trained model of the CaffeNet version [Ji14] of AlexNet trained on the ImageNet dataset. Following the architecture of AlexNet, the input is an image of size 227×227, randomly cropped from an image of size 256×256, to make the network robust to changes in translation. Both colour and depth images are therefore resized accordingly before being processed by the network. In addition, the images are randomly flipped to increase the amount of training data. In case of depth images, [Ei15] shows that applying a jet colourmap enhances the accuracy compared to encoding the images using surface normals [BRF13] or Horizontal disparity, Height and Angle (HHA) encoding [Gu14]. This colour transformation maps each depth value to a colour in RGB colour space from blue(close) over green to red(far). This enables us to initialize the weights using the pre-trained CaffeNet model without additional preprocessing. We therefore perform similar step before training the depth model.

Given sets of parameters and datasets $(W^{RGB}, b^{RGB}, X^{RGB}, Y)$ and (W^{D}, b^{D}, X^{D}, Y) for RGB and depth, respectively, where W and b are the model weights and bias, while (X^{RGB}, X^{D}) are the set of RGB and depth images with corresponding labels Y, we train the models by minimizing a loss function, L, as given in Equation 1:

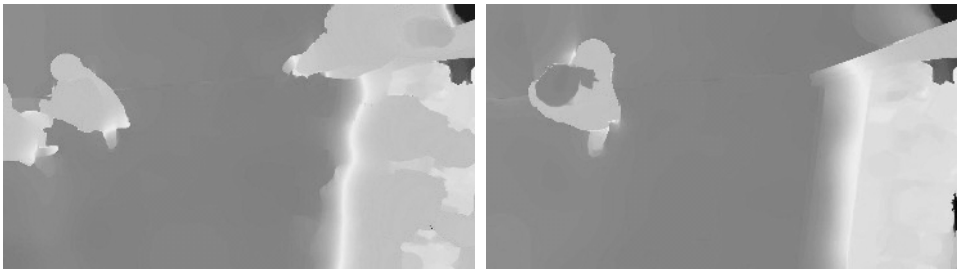
$$\min_{W,b} -\frac{1}{N}\sum_{i=1}^{N} L(W,b,x_i,y_i) \qquad L(W,b,x_i,y_i) = log(\hat{p}_i, y_i) \tag{1}$$

where W, b are the weights and bias of the model currently being trained, $X = \{x_1, .., x_N\}$ is the sample set and $\hat{p}_i$ is the output probability from the softmax layer of the i'th sample given the true label y_i.

Joint training After training RGB-CNN and D-CNN, the model parameters are used to initialize the two CNN streams in RGB-D-CNN. The softmax layers are replaced by a randomly initialized fully connected layer (fc8) and new softmax layer (fc9). By fusing outputs from both fc7^{RGB} and fc7^{D} in fc8, the parameters of the depth stream are updated depending on the input to the RGB stream and vice versa, while the weights and bias of fc8 are updated based on both inputs, resulting in a fused output. [WCZ16] shows how fusion of hand-crafted and CNN features in the late layers of the network affects parameter update of the CNN. Similar proof applies to this context.

4 Experimental Results

Datasets For evaluation we present a novel RGB-D based dataset collected from an overhead view. We refer to the dataset as Overhead Person Re-identification (OPR). The dataset is collected using a calibrated ZED stereo camera from Stereolabs [St17], mainly due to its ability to record depth from a range 0.7m-20m covering both low and high ceilings. In addition, it captures video in resolutions up to 4416×1242 pixels which is much higher than RGB sensors in solutions such as the Microsoft Kinect. The camera is placed in the ceiling at a university canteen (uncontrolled environment) to capture a populated area. From this perspective, persons are captured when approaching (walking from top to bottom), and leaving (walking from bottom to top) the canteen a few minutes later, enabling us to evaluate re-id performance. Data is collected on a single day during a two hour period around midday to capture video when the number of persons in the canteen is increasing and decreasing. As a result, cases of having a large number of persons and only a single person are recorded, examples of captured depth images in both cases are shown in Figure 2 (a) and (b), respectively.



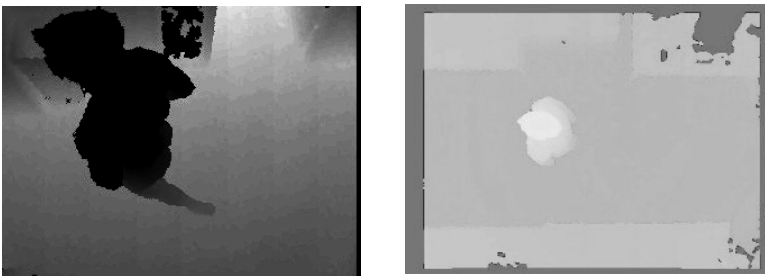
(a) (b)

Fig. 2: Examples of depth images containing (a): multiple persons and (b): containing a single person. Each person is captured when approaching (right side) and leaving (left side) the canteen.

Disparity maps are computed using Semi-Global Block Matching (SGBM) as it has shown as a good compromise between accuracy and processing time [Ka11], followed by filtering

using a Weighted Least Square (WLS) kernel to eliminate noise and make the background more uniform, resulting in more precise depth information. Finally, we manually annotate bounding boxes around persons and use those for our system, the annotations enables us to further test detection, tracking and segmentation algorithms in future work. A total number of 78742 frames with 64 different persons have been annotated for re-id.

To our knowledge, only the datasets of [HAFF16] (DPI-T) and [Li17] (TVPR) have previously been proposed for RGB-D and overhead based re-id. Both are recorded in a hall with only a single person within view at all times. Examples of depth images from these datasets are shown in Figure 3. In addition to evaluating on our own dataset, we apply our system to those of [HAFF16] and [Li17] for comparison with their original results.



(a) (b)

Fig. 3: Examples of depth images from (a): DPI-T and (b): TVPR.

Evaluation protocols Depending on the dataset, different training and testing protocols are followed.

OPR Similar to most RGB-based datasets within re-id, we perform 10 random train and test splits, each set containing 32 persons. After training the CNN models, features from the test set are extracted from the last fully connected layer.

TVPR The training set originally consists of 100 persons walking from left to right while the test set consists of same persons walking from right to left. During test, features from the test set are compared with those from the training set. Although, due to issues at test time regarding one of the video sequences, only 94 persons were considered for training and testing.

DPI-T 12 persons appear in five different sets of clothing in both the training and test, while the number of recordings in each set differs. A total of 213 sequences are used for training while the test set consists of 249 sequences which are all classified by comparing with those of the training set.

When training RGB-CNN and D-CNN, a batch size of 128 is used while a size of 64 is used in case of RGB-D-CNN. Network parameters are updated using Stochastic Gradient Descent (SGD) with momentum is to avoid getting stuck in a local minimum. Hyper parameters are set accordingly to [Ei15] with a momentum of 0.9 and base learning rate of 0.01 which is reduced by multiplying with 0.97 for each epoch. At each epoch, the training set is randomly shuffled for faster convergence [Be12]. We present our results by calculating the *rank-1* to *rank-k* accuracies based on feature matching where *rank-i* indicates a

cumulative percentage of persons having their true match within the i most similar with k indicating the total number of persons. For OPR, the average accuracies over all train/test splits are calculated. Matches are calculated using Euclidean distance between extracted features following a multi-shot approach, i.e. features from all images of each person/sequence are extracted and either maximized or averaged, indicated by subscripts *max* and *avg*.

Figure 4 (a) shows the resulting Cumulative Matching Characteristic (CMC) curves for applying RGB-CNN, D-CNN and RGB-D-CNN to OPR. It is clear that fusing of RGB and depth modalities clearly increases accuracy compared to using a single modality. The best result is achieved by RGB-D-CNN$_{avg}$, increasing accuracy by 16.00% compared to RGB-CNN$_{avg}$. Furthermore, Figure 4 (b) and (c) show the results of our system applied to DPI-T and TVPR, respectively. In case of DPI-T, RGB-D-CNN$_{avg}$ still outperforms RGB-CNN and D-CNN with an increase of 3.61% compared to RGB-CNN$_{avg}$. Finally for TVPR, RGB-CNN provide better results compared to RGB-D-CNN. A reason for this could be the quality of depth information (see Figure 3 (b)) negatively affecting the training of RGB-D-CNN in combination with corresponding colour images. Even though, D-CNN results are slightly worse in case of DPI-T, the level of detail in depth images are higher (see 3 (a)) causing the modality to better complement RGB. The quality of depth information therefore seems important when training an RGB-D CNN. Looking at results across all datasets, averaged features mostly provides the highest accuracies, although, in case of depth features, feature maximization seems better. This could be due to encoding of features as colourized images combined with an overhead view from which the height of each person, and thereby the colour gradient, is important. By averaging features, this information more easily gets lost if the representation changes between images.

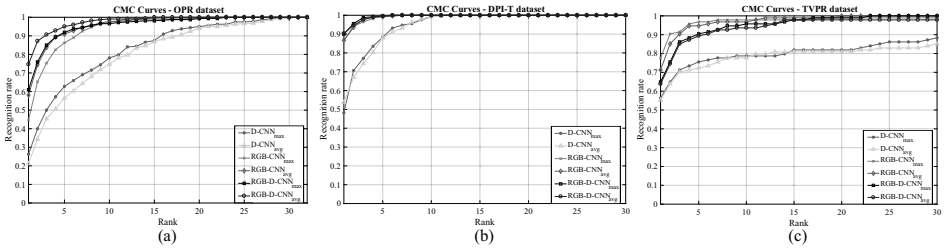


Fig. 4: Results on (a) OPR (p=32), (b) DPI-T (p=249) and (c) TVPR (p=94) for RGB-CNN, D-CNN and RGB-D-CNN, respectively, using maximized (*max*) and averaged (*avg*) features.

Tables 1 summarizes our results on TVPR and DPI-T, compared to their original results. As [HAFF16] only provides a rank-1 accuracy while [Li17] only provides CMC curves, only the rank-1 accuracy is considered. For [Li17], rank-1 is estimated from the CMC curves. *Ours* refers to the best results achieved by our system (RGB-D-CNN$_{avg}$ in case of DPI-T and RGB-CNN$_{avg}$ in case of TVPR). In both cases we outperform original results, for DPI-T by 34.76% by also using RGB. From Figure 4 (b), it is worth noting that our D-CNN alone achieves almost similar accuracies as [HAFF16] who also adds an LSTM layer on top of a similar CNN.
Even though, six persons are missing for the tests on TVPR, our system shows potential to be improved further. For RGB alone, our system outperforms that of [Li17] by ≈5.16%.

	Rank-1 accuracy [%]	
Method	*DPI-T*	*TVPR*
4D RAM [HAFF16]	55.60	–
TVDH [Li17]	–	72.50
Ours	**90.36**	**77.66**

Tab. 1: Comparison of our RGB-D-CNN to original results on DPI-T and TVPR datasets.

Processing time We evaluate processing time for stereo and feature matching on OPR to discuss on the potential of using passive stereo for re-id applications. 20 matching iteration are run using an Intel i7-6700HQ CPU @ 2.60GHz and 16GB of RAM and average timings are provided. Stereo matching matching is performed on images of size 960×540.

While feature matching only takes $4.0e10^{-5}$s, SGBM and WLS are more processing intensive taking 0.136s and 0.103s, respectively. Nonetheless, $\approx$4 FPS is achieved using the CPU. For real-time applications, GPU implementations of SGBM and WLS algorithms could be used speed up the process. No such implementations are available at the moment.

5 Conclusion

In this paper, we have presented an RGB-D based CNN applied to person re-identification. Two CNN models are trained using colour and depth images, respectively, captured from an overhead view and resulting trained parameters are used to initialize a joint RGB-D-CNN model trained using both modalities. To test the system, we collected a novel RGB-D and overhead based dataset which is annotated for evaluation on both re-id accuracy, but also detection and tracking algorithms. By applying our system to our novel and two previously proposed datasets, we have shown that the combination of RGB and depth modalities increase accuracy by 16.0% and 3.6% on our OPR dataset and DPI-T, respectively. In case of TVPR, RGB modality alone achieved higher accuracy than combining modalities due to the quality of depth information. This indicates an importance to capture detailed depth information to proper complement the RGB modality. In addition, our system shows an FPS of 4 using a CPU, with potential of being increased if processing intensive algorithms such as SGBM and WLS are implemented on a GPU. For future work, the system should be evaluated on bounding boxes extracted automatically from a person detector. To increase detection performance, depth information could also be used for this purpose. Furthermore, our proposed system could be extended with an LSTM to handle video rather than averaging or maximizing features extracted from a sequence of images. This would allow for temporal information to be captured as well. Finally, more recently developed neural networks could replace the AlexNet architecture to increase performance and decrease processing time.

Acknowledgement

The work carried out in this paper is supported by Innovation Fund Denmark under Grant 5189-00222B.

References

[AC12] Ahmed, Imran; Carter, John N: A robust person detector for overhead views. In: Proc. ICPR. IEEE, pp. 1483–1486, 2012.

[AJM15] Ahmed, Ejaz; Jones, Michael; Marks, Tim K: An improved deep learning architecture for person re-identification. In: Proc. CVPR. IEEE, pp. 3908–3916, 2015.

[Ar08] Aradhye, Hrishikesh; Fischler, Martin; Bolles, Robert; Myers, Gregory: Headprint-Based Human Recognition. In: Advances in Biometrics: Sensors, Algorithms and Systems. Springer London, chapter 15, pp. 287–306, 2008.

[Ba12] Barbosa, Igor; Cristani, Marco; Del Bue, Alessio; Bazzani, Loris; Murino, Vittorio: Re-identification with rgb-d sensors. In: Computer Vision–ECCV 2012: Workshops and Demonstrations. Springer, pp. 433–442, 2012.

[Be12] Bengio, Yoshua: Practical recommendations for gradient-based training of deep architectures. In: Neural networks: Tricks of the trade, pp. 437–478. Springer, 2012.

[BRF13] Bo, Liefeng; Ren, Xiaofeng; Fox, Dieter: Unsupervised feature learning for RGB-D based object recognition. In: Experimental Robotics. Springer, pp. 387–402, 2013.

[Ch16] Chen, Ying-Cong; Zheng, Wei-Shi; Lai, Jian-Huang; Yuen, Pong: An Asymmetric Distance Model for Cross-view Feature Mapping in Person Re-identification. IEEE Transactions on Circuits and Systems, 2016.

[Ei15] Eitel, Andreas; Springenberg, Jost Tobias; Spinello, Luciano; Riedmiller, Martin; Burgard, Wolfram: Multimodal deep learning for robust rgb-d object recognition. In: Proc. IROS. IEEE, pp. 681–687, 2015.

[GT08] Gray, Douglas; Tao, Hai: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proc. ECCV. Springer, pp. 262–275, 2008.

[Gu14] Gupta, Saurabh; Girshick, Ross; Arbeláez, Pablo; Malik, Jitendra: Learning rich features from RGB-D images for object detection and segmentation. In: Proc. ECCV. Springer, pp. 345–360, 2014.

[Ha16] Hazirbas, Caner; Ma, Lingni; Domokos, Csaba; Cremers, Daniel: Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian Conference on Computer Vision. Springer, pp. 213–228, 2016.

[HAFF16] Haque, Albert; Alahi, Alexandre; Fei-Fei, Li: Recurrent Attention Models for Depth-Based Person Identification. In: Proc. CVPR. IEEE, pp. 1229–1238, 2016.

[Ji14] Jia, Yangqing; Shelhamer, Evan; Donahue, Jeff; Karayev, Sergey; Long, Jonathan; Girshick, Ross; Guadarrama, Sergio; Darrell, Trevor: Caffe: Convolutional architecture for fast feature embedding. In: Proc. ACMMM. ACM, pp. 675–678, 2014.

[Ka11] Kalarot, Ratheesh; Morris, John; Berry, David; Dunning, James: Analysis of real-time stereo vision algorithms on GPU. In: Proc. IVCNZ. 2011.

[Ki98] Kittler, Josef; Hatef, Mohamad; Duin, Robert PW; Matas, Jiri: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):226–239, 1998.

[KSH12] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. Citeseerx, pp. 1097–1105, 2012.

[Li14] Li, Wei; Zhao, Rui; Xiao, Tong; Wang, Xiaogang: Deepreid: Deep filter pairing neural network for person re-identification. In: Proc. CVPR. IEEE, pp. 152–159, 2014.

[Li15] Liao, Shengcai; Hu, Yang; Zhu, Xiangyu; Li, Stan Z: Person re-identification by local maximal occurrence representation and metric learning. In: Proc. CVPR. IEEE, pp. 2197–2206, 2015.

[Li17] Liciotti, Daniele; Paolanti, Marina; Frontoni, Emanuele; Mancini, Adriano; Zingaretti, Primo: Person Re-identification Dataset with RGB-D Camera in a Top-View Configuration. In: Video Analytics. Face and Facial Expression Recognition and Audience Measurement. Springer, pp. 1–11, 2017.

[LSF15] Li, Sheng; Shao, Ming; Fu, Yun: Cross-view projective dictionary learning for person re-identification. In: Proc. AAAI. AAAI Press, pp. 2155–2161, 2015.

[Mu14] Munaro, Matteo; Basso, Alberto; Fossati, Andrea; Van Gool, Luc; Menegatti, Emanuele: 3D reconstruction of freely moving persons for re-identification with a depth sensor. In: Proc. ICRA. IEEE, pp. 4512–4519, 2014.

[Sa16] Sanchez-Riera, Jordi; Hua, Kai-Lung; Hsiao, Yuan-Sheng; Lim, Tekoing; Hidayati, Shintami C; Cheng, Wen-Huang: A comparative study of data fusion for RGB-D based visual recognition. Pattern Recognition Letters, 73:1–6, 2016.

[St17] Stereolabs: , ZED - Depth Sensing and Camera Tracking. https://www.stereolabs.com/zed/specs/, 2017.

[WCZ16] Wu, Shangxuan; Chen, Ying-Cong; Zheng, Wei-Shi: An enhanced deep feature representation for person re-identification. In: Proc. WACV. IEEE, pp. 1–8, 2016.

[ZXG16] Zhang, Li; Xiang, Tao; Gong, Shaogang: Learning a Discriminative Null Space for Person Re-Identification. In: Proc. CVPR. IEEE, pp. 1239–1248, 2016.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 35

Pool Adjacent Violators Based Biometric Rank Level Fusion

Nanang Susyanto[1]

Abstract: We propose a new method in rank level fusion for biometric identification. Our method is based on the pool adjacent violators (PAV) algorithm after the ranks have been transformed to the approximated scores. We then show that our method outperforms various approaches that commonly used in biometric rank level fusion on NIST BSSR1 multimodal database.

Keywords: rank level fusion, pool adjacent violators, biometric fusion, biometric identification, multimodal.

1 Introduction

Biometric fusion is a combination of several biometric systems or algorithms that aims to improve the performance of the individual system or algorithm. It can be divided into six categories [RNJ06]: multi-sensor, multi-algorithm, multi-instance, multi-sample, multi-modal and hybrid. Several studies show the performance improvement by combining information from multiple traits or algorithms [LWJ03, RJR02, RNJ06, Ul06]. For instance, Lu et al. [LWJ03] combine three different feature extractions (Principle Component Analysis, Independent Component Analysis and Linear Discriminant Analysis) while Prabhakar and Jain [PJ02] in the fingerprint biometric field use the left and right index fingers to verify an individual's identity.

Biometric fusion can be done at the sensor, feature, match score, rank and decision levels either for verification or identification. In this paper, we will focus on the rank level for closed identification problem in the sense that the unknown person is assumed to be one of the people in a given enrollment set. This scenario is suitable for combining ranked identities from commercial biometric devices that may only produce the ranked identities of the users instead of matching scores because of a security reason. This means that ranks from multiple biometric classifiers of every unknown person in a given enrollment set are transformed to a new rank and this new rank is used to assign the identity of that unknown person.

There are several methods that are commonly used in biometric rank level fusion: Borda count, weighted Borda count, maximum rank method, Bucklin majority voting [Po14], and some nonlinear weighted ranks [KS11]. While maximum rank takes the highest rank amongst all the matchers, the remaining methods use weight to represent the contribution of each classifier. The present paper uses the pool adjacent violators (PAV) algorithm to

[1] Faculty of Mathematics and Natural Sciences (FMNS), Department of Mathematics, Universitas Gadjah Mada, Sekip Utara Yogyakarta, Indonesia, nanang_susyanto@ugm.ac.id

compute the likelihood ratio (LR) of any rank after it has been transformed to its approximated similarity score for every classifier and combine the classifiers by summing their individual LRs up to get the final score. This final score will represent the combined similarity score. The rest of this paper is organized as follows. Section 2 gives a detailed explanation how the proposed method woks. Several examples using NIST BSSR1 database are provided in Section 3. Finally, this paper will be closed by our conclusions in Section 4.

2 PAV-based Method

This section will explain how our proposed method is built. In principle, there are two steps: (1) transforming ranks to their approximated similarity scores and (2) applying the PAV to these transformed scores.

2.1 Transforming Ranks to Approximated Scores

Let x be an unknown subject that belongs to the enrollment set $E = \{x_1, \ldots, x_n\}$. Of course, the original similarity scores of x and all elements E contain much more information than the ranked identity of all elements E with respect to the closeness to x. While Susyanto et al. [Su16a, Su16b] use a modified empirical distribution function to transform similarity scores to their corresponding uniformly distributed scores, which are only a scale of their ranks, to model dependence between classifiers, we will work on the other direction, i.e, approximating the uniformly distributed similarity scores from their ranks. Suppose that there are n_{train} identities in the enrollment set in *training* data. Since the rank-i has to have the i-highest probability for every $i = 1, \ldots, n_{\text{train}}$, we set it to have probability $(n_{\text{train}} + 1 - i)/n_{\text{train}}$. It means that the estimated probabilities run from $1/n_{\text{train}}$ to 1, which is already shown in [Su16a, Su16b] that they are uniformly distributed. Below is an example how the approximated similarity scores of the training set with subjects s_1, s_2, s_3, s_4, and s_5 are obtained from the ranks.

$$\begin{pmatrix} enr. & s_1 & s_2 & s_3 & s_4 & s_5 \\ s_1 & 1 & 2 & 3 & 4 & 5 \\ s_2 & 1 & 4 & 3 & 5 & 2 \\ s_3 & 2 & 5 & 1 & 3 & 4 \\ s_4 & 5 & 4 & 3 & 2 & 1 \\ s_5 & 5 & 2 & 3 & 4 & 1 \end{pmatrix} \mapsto \begin{pmatrix} enr. & s_1 & s_2 & s_3 & s_4 & s_5 \\ s_1 & 1 & 4/5 & 3/5 & 2/5 & 1/5 \\ s_2 & 1 & 2/5 & 3/5 & 1/5 & 4/5 \\ s_3 & 4/5 & 1/5 & 1 & 3/5 & 2/5 \\ s_4 & 1/5 & 2/5 & 3/5 & 4/5 & 1 \\ s_5 & 1/5 & 4/5 & 3/5 & 2/5 & 1 \end{pmatrix} \tag{1}$$

When we are working in a testing data that contains more that n_{train} in its enrollment set, then we map all ranks greater than n_{train} to 0. Mathematically, whenever the training data T with n_{train} identities in its enrollment set is given and the subjects $x_1, \ldots, x_n$ has rank $r_1, \ldots, r_n$, respectively, with respect to the unknown subject x, the approximation of the uniformly distributed scores $s_1, \ldots, s_n$ will be

$$s_i = \max\left\{\frac{n_{\text{train}} + 1 - i}{n_{\text{train}}}, 0\right\} \tag{2}$$

for every $i = 1, \ldots, n$. For example, if $x_1, x_2, x_3, x_4, x_5, x_6$, and x_7 in the testing set has ranks $2, 3, 1, 4, 5, 6$, and 7, respectively, with respect to x_3 then the approximated similarity scores using training data (1) is

$$\begin{pmatrix} enr. & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ x_3 & 4/5 & 3/5 & 1 & 2/5 & 1/5 & 0 & 0 \end{pmatrix}. \tag{3}$$

2.2 PAV-based Naive Bayes Fusion

Once we have had the approximated similarity scores of the training data, we can split them into *genuine* and *impostor* scores. A genuine score is the score obtained by comparing a pair of biometric samples originating from the same person while an impostor scores obtained by comparing a pair of biometric samples stemming from different people. In our example given by (1), the elements on the main diagonal of the second matrix are genuine while the elements off the main diagonal are impostor. The key of our method is computing the loglikelihood (LLR) of the approximated similarity scores using independence assumption. Even though it is not realistic, some experimental results show that its performance is still promising [TV13]. To do that, we need to compute the individual LLRs of the classifiers. The most common methods are Kernel Density Estimation (KDE), Logistic Regression (Logit), Histogram Binning (HB), and Pool Adjacent Violators (PAV); see [ASV12] for a brief explanation of these methods.

In this paper, we choose the PAV method because of its optimality [ZE02]. For every classifier $k = 1, \cdots, d$, (d is the number of classifiers), PAV sorts and assigns a posterior probability of 1 and 0 to the k-th component of genuine and impostor scores, respectively, in a given training set. It then computes the non-monotonic adjacent group of probabilities and replaces it with average of that group. This procedure is repeated until the whole sequence is monotonically increasing which estimates the posterior probability $P(H_1|(\cdot))$ of the k-th component of genuine and impostor scores where H_1 correspond to a genuine score. By assuming

$$P(H_1) = \frac{n_{\text{gen}}}{n_{\text{gen}} + n_{\text{imp}}},$$

the corresponding LR_ks of genuine and impostor scores can be computed according to the Bayesian formula by

$$\widehat{\text{LR}}_k(\cdot) = \frac{P(H_1|(\cdot))}{1 - P(H_1|(\cdot))} \times \frac{n_{\text{imp}}}{n_{\text{gen}}} \tag{4}$$

where n_{gen} and n_{imp} are the number of genuine and impostor scores, respectively. This gives a numerical function that maps score to its $\widehat{\text{LR}}_k$ so that for every score from the k-th classifier, its corresponding $\widehat{\text{LR}}_k$ value can be computed by interpolation. Finally, the final approximated similarity score is just the sum of all $\widehat{\text{LR}}_k$s for $k = 1, \ldots, d$.

3 Experimental Results

This section gives the comparison between the proposed method and the existing methods in rank level fusion (Borda count, weighted Borda count, maximum rank method, Bucklin majority voting [Po14], and some nonlinear weighted ranks [KS11]) on NIST BSSR1 database [Na04]. The NIST-BSSR1 database has three different set:

- NIST-Multimodal: Two fingerprints and Two face matchers applied to 517 subjects,
- NIST-Face: Two face matchers applied to 3000 subjects,
- NIST-Finger: Two fingerprints applied to 6000 subjects.

We will use the same protocol as used in [KS11] (Protocol 1 and 2) and an additional protocol (Protocol 3). The exp(1) and exp(2) are the methods proposed in [KS11]. Note that the maximum rank, the Borda count, and the Bucklin majority voting methods do not need training data while our proposed method needs training data as the weighted Borda count and nonlinear weighted ranks [KS11]) do.

3.1 Multi-instance Test: Protocol 1

In this experiment, we use the NIST-Finger database containing 6000 subject where the first 1000 subject were used for training our proposed method and the rest were used for testing. The comparison of our method with the other methods is presented in Table 1. We can see that the rank-1 of our method jumps from the best existing methods (exp(2): 89.56%) to 94.44%.

Tab. 1: Performance (in %) From NIST-Finger Database (6000 Subjects). The bold face in every row is the best one.

	Highest Rank	Borda	Weighted Borda	Bucklin	exp(1)	exp(2)	Proposed
rank-1	82.57	85.65	87.74	74.58	89.34	89.56	**94.44**
rank-2	94.43	86.68	89.04	88.23	93.98	94.42	**95.00**
rank-3	94.48	87.33	89.74	93.65	95.22	95.20	**95.34**

3.2 Multi-modal and Multi-algorithm Test: Protocol 2

Using the same protocol as in [KS11], we put the first 100 subject of the NIST-Multimodal database for training and the rest (417) for testing. We can see from Table 2 that our proposed method outperforms the other methods even it attains 100% recognition rate at rank-1.

Tab. 2: Performance (in %) From NIST-Multimodal Database (517 Subjects). The bold face in every row is the best one.

	Highest Rank	Borda	Weighted Borda	Bucklin	exp(1)	exp(2)	Proposed
rank-1	80.66	91.68	94.39	88.78	98.84	99.28	**100.00**
rank-2	96.32	93.81	95.55	98.84	99.42	99.76	**100.00**
rank-3	100.00	94.97	96.32	99.81	100.00	100.00	**100.00**

3.3 Multi-modal and Multi-algorithm Test: Protocol 3

In order to make a larger database for testing, we make a virtual database by taking the first image of every person in NIST-Face database and the fist 3000 subjects in NIST-Finger. As the results, our virtual database contains 3000 subjects in which every subject has 2 scores from face comparisons, 1 score from left-index finger comparison, and 1 score from right-index finger comparison. By using the same training data as in Protocol 2, we can see from Table 2 that the highest rank, Borda count, weighted Borda count, and Bucklin methods do not perform better than the exp(1) and exp(2) methods. Therefore, we will only compare our method with the exp(1) and exp(2) methods. The cumulative match curve is provided by Figure 1 thats shows clearly that the proposed method does outperform the exp(1) and exp(2). The recognition rate at rank-1 of the exp(1), exp(2), our proposed method are 95.73%, 91.97%, and 98.87%, respectively.

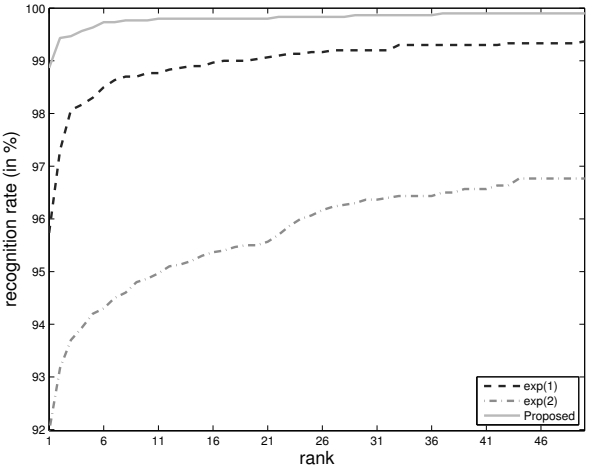


Fig. 1: CMC of the exp(1), exp(2), and the Proposed Method

4 Conclusion

We have proposed a new method in biometric rank level fusion via pool adjacent violators (PAV). The method can be done by two main steps: (1) transforming ranks to their approxi-

mation of the uniformly distributed similarity scores and (2) applying the PAV of the transformed scores for every classifier and simply taking the naive Bayes fusion. It has been demonstrated that our proposed method outperforms the Borda count, weighted Borda count, maximum rank method, Bucklin majority voting, and some nonlinear weighted ranks in every scenario using the NIST BSSR1 database.

References

[ASV12] Ali, T.; Spreeuwers, L. J.; Veldhuis, R. N. J.: Forensic Face Recognition: A Survey. In (Quaglia, A.; Epifano, C. M., eds): Face Recognition: Methods, Applications and Technology, Computer Science, Technology and Applications, p. 9. Nova Publishers, 2012.

[KS11] Kumar, A.; Shekhar, S.: Personal Identification Using Multibiometrics Rank-Level Fusion. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41(5):743–752, Sept 2011.

[LWJ03] Lu, Xiaoguang; Wang, Yunhong; Jain, A.K.: Combining classifiers for face recognition. In: Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on. volume 3, pp. III–13–16 vol.3, July 2003.

[Na04] National Institute of Standards and Technology: , NIST Biometric Scores Set - release 1, 2004. Available at http://www.itl.nist.gov/iad/894.03/biometricscores.

[PJ02] Prabhakar, Salil; Jain, Anil K.: Decision-level fusion in fingerprint verification. Pattern Recognition, 35(4):861 – 874, 2002.

[Po14] Porter, Melvin P.: Preferential voting and the rule of the majority. National Municipal Review, 3(3):581–585, 1914.

[RJR02] Ross, A.; Jain, A.; Reisman, J.: A hybrid fingerprint matcher. In: Pattern Recognition, 2002. Proceedings. 16th International Conference on. volume 3, pp. 795–798 vol.3, 2002.

[RNJ06] Ross, Arun; Nandakumar, Karthik; Jain, Anil K.: Handbook of Multibiometrics (International Series on Biometrics). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[Su16a] Susyanto, N.; Veldhuis, R. N. J.; Spreeuwers, L. J.; Klaassen, C. A. J.: Fixed FAR Correction Factor of Score Level Fusion. In: The 8th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS2016), 6-9 September, Buffalo, NY, USA. September 2016.

[Su16b] Susyanto, N.; Veldhuis, R. N. J.; Spreeuwers, L. J.; Klaassen, C. A. J.: Two-step Calibration Method for Multi-algorithm Score-based Face Recognition Systems by Minimizing Discrimination Loss. In: The 9th IAPR International Conference on Biometrics (ICB 2016), 13-16 June, Halmstad. June 2016.

[TV13] Tao, Q.; Veldhuis, R. N. J.: Robust Biometric Score Fusion by Naive Likelihood Ratio via Receiver Operating Characteristics. IEEE Transactions on Information Forensics and Security, 8(2):305–313, February 2013.

[Ul06] Ulery, Brad; Hicklin, Austin; Watson, Craig; Fellner, William; Hallinan, Peter; Gutierrez, Carlos M: Studies of Biometric Fusion NISTIR 7346. 2006.

[ZE02] Zadrozny, Bianca; Elkan, Charles: Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '02, ACM, New York, NY, USA, pp. 694–699, 2002.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 41

Improvement of Iris Recognition based on Iris-Code Bit-Error Pattern Analysis

Christian Rathgeb[1], Christoph Busch[1]

Abstract: In this paper an advanced iris-biometric comparator is presented. In the proposed scheme an analysis of bit-error patterns produced by Hamming distance-based iris-code comparisons is performed. The lengths of sequences of horizontal consecutive mis-matching bits are measured and a frequency distribution is estimated. The difference of the extracted frequency distribution to that of an average genuine one obtained from a training set is used as a second comparison score. This score is then used together with the fractional Hamming distance in order to improve the recognition accuracy of an iris recognition system. In experimental evaluations relative improvements of approximately 45% and 10% in terms of false non-match rate at a false match rate of 0.01% are achieved on the CASIAv4-Interval and the BioSecure iris databases, respectively.

Keywords: Biometrics, iris recognition, iris-code, bit-error analysis, improved comparator.

1 Introduction

Generic iris recognition systems comprise four major components: (1) image acquisition, (2) segmentation, (3) feature extraction and (4) comparison. Based on Daugman's approach [Da04], the first three processing steps are performed on a reference iris image during enrolment to create a two-dimensional binary feature vector, i.e. iris-code. At the time of authentication an iris-code is extracted from a probe iris image and compared against a database of enrolled reference iris-codes. In the comparison stage Hamming distance (*HD*) scores between pairs of iris-codes and corresponding noise masks are estimated. Hence, the binary data representation of iris-codes enables a rapid comparison (and compact storage) achieving millions of comparisons per second per CPU core [Da04]. Circular bit shifts are applied to iris-codes and *HD* scores are estimated at different shifting positions, i.e. relative tilt angles caused by uncontrolled head poses. The minimal obtained *HD*, which corresponds to an optimal alignment, represents the final score.

Besides the Daugman de-facto standard for comparing iris-codes, different alternative comparators have been suggested in past years, see Sect. 2. The majority of proposed schemes aims at replacing the aforementioned *HD*-based algorithm by a modified comparator in order to improve the recognition performance. In most schemes findings obtained from a deeper analyses of the nature of the iris-code bits are utilized by those comparators. A prominent example for such an improvement is the assignment of weights to each bit position in an iris-code according to their expected *reliability*, e.g. in [ZD08, DST11].

[1] da/sec – Biometrics and Internet Security Research Group,
Hochschule Darmstadt, Germany, {christian.rathgeb,christoph.busch}@h-da.de

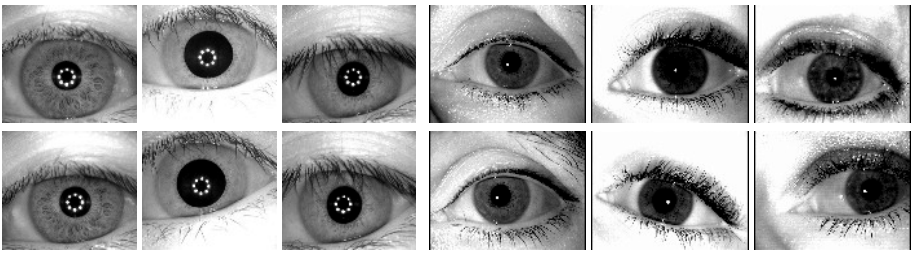
In this work we analyse entire bit-error patterns produced by *HD*-based iris-code comparisons, going beyond a local estimation of bit-errors. The presented approach measures the plausibility of an obtained bit-error pattern by comparing it to a pre-estimated model of genuine bit-error patterns. In particular, the frequency distribution of sequences of horizontal, i.e. circumferential, consecutive mis-matching bits is measured and its difference from the genuine model is used as secondary feature. This score can be interpreted as additional score, which can be estimated to achieve a more reliable decision for a distinct range of *HD* scores, e.g. $[0.35, 0.45]$. Hence, in contrast to most proposed comparators, our approach is designed to have negligible impact on comparison speed. For different iris databases it is shown that a weighted score-level fusion of the proposed score and the *HD* score improves the recognition accuracy of an iris recognition system, in particular at practical low false match rates.

The remainder of this paper is organized as follows: Sect. 2 briefly summarizes related works with respect to iris-biometric comparators. In Sect. 3 the proposed system is described in detail and evaluated. Finally, conclusions are drawn and potential future research directions are pointed out in Sect. 4.

2 Related Works

In the recent past numerous improved iris-biometric comparators have been proposed. Some of these require the processing of multiple reference samples during enrolment. In [ZD08] a weight map, which indicates the stability of iris-code bits, is obtained from several iris-codes by performing a weighted majority voting. Similar approaches based on personalized weight maps have been presented in [DST11, HSH17]. In these schemes comparison scores are estimated as a weighted sum of mis-matching bits. Note that for these modified comparators one can not expect that the comparison speed of a Hamming distance-based comparator is maintained. In [HBF09] so-called *fragile* bits, i.e. bits which exhibit a higher probability than others to flip their value during a genuine comparison, are detected by comparing several iris-codes obtained from a single eye instance. Since filters employed in the feature extraction stage set iris-code bits by the sign of obtained filter responses, these bits correspond to coefficients close to zero. That is, such bits can also be detected in a single iris-code [Da16]. It was shown that the recognition accuracy is improved, if detected fragile bits are incorporated into noise masks extracted in the iris segmentation stage. Moreover, masks encoding fragile bits can be employed as additional comparison sore to improve the performance of an iris recognition system [HBF11].

Further works utilize training sets to obtain statistics about iris-codes which are utilized by the comparator. In [RUW10] a static weight map indicating the reliability of each iris-code bit position, which is defined as the mean of discriminativity and stability, is estimated from a training set. During authentication most reliable bits are compared first to achieve a fast rejection of non-matching iris-codes in an identification scenario. A similar approach based on static masks has been presented in [Pr15]. Reported results suggest that static weight maps might vary depending on the used sensor or environmental conditions. In [RUW12] the progression of genuine comparison scores across all considered shifting



(a) CASIAv4-Interval (b) BioSecure

Fig. 1: Sample pairs of iris images of both datasets used in experimental evaluations.

Tab. 1: Overview of training and testing sets of employed datasets.

Database	Training set (left eye images)			Testing set (right eye images)		
	No. eyes	Gen. comp.	Imp. comp.	No. eyes	Gen. comp.	Imp. comp.
CASIAv4-Interval	198	4,454	19,503	197	4,343	19,306
BioSecure	210	1,260	21,945	210	1,260	21,945

positions are modelled by an inverse Gaussian of which the parameters are estimated from a training set. At authentication the deviation of comparison scores from the trained Gaussian is combined with the minimum *HD* score.

Given a single pair of iris-codes, in [RUW11] it is suggested to combine the minimum and the maximum *HD* score across shifting positions. Since genuine pairs of iris-codes can get out of phase in case of drastic mis-alignment exceptionally large *HD* scores become an indicator for a genuine comparison. More recently, a binary search technique which aims at accelerating the alignment process during iris-code comparisons was presented in [Ra16]. It is shown that, if the amount of considered shifting positions can be reduced, recognition accuracy is generally improved since *HD* scores of impostor comparisons remain higher.

3 Proposed System

3.1 Baseline System and Experimental Setup

In the employed iris recognition system, the iris of a given sample image is detected and transformed to a normalized rectangular texture of 512×64 pixels. The normalized iris texture is divided into texture stripes to obtain 10 one-dimensional signals, each one averaged from adjacent texture rows. A row-wise convolution with a Log-Gabor wavelet is performed on each signal and the real part of phase information is encoded to generate an iris-code consisting of 512×10 bits. Examples of generated iris-codes are depicted in Fig. 2. Implementations of the employed segmentation and feature extraction are available in [US17] and described in detail in [RUW13].

The fractional Hamming distance (*HD*) between a pair of iris-codes, *codeA*, *codeB*, and their according noise masks, *maskA*, *maskB* is defined as [Da04],

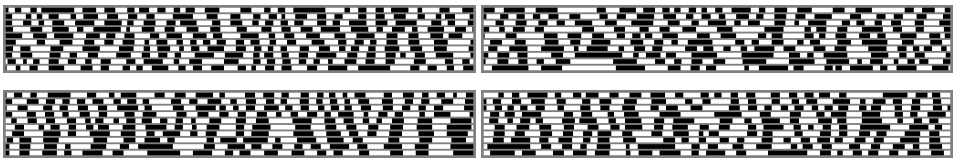
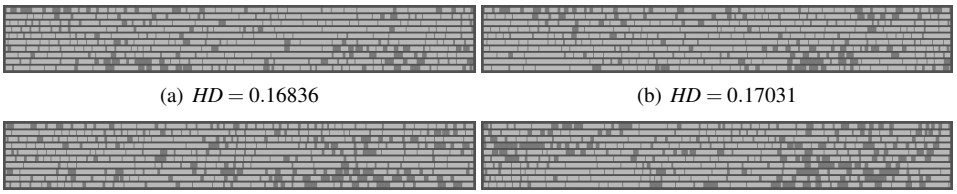


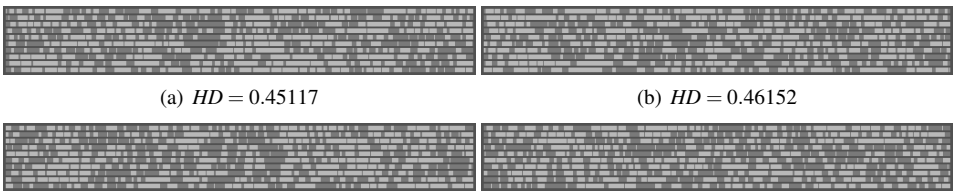
Fig. 2: Examples of iris-codes produced by four different iris images of used datasets.



(a) $HD = 0.16836$ (b) $HD = 0.17031$

(c) $HD = 0.20957$ (d) $HD = 0.24316$

Fig. 3: Examples of bit-error patterns produced by four genuine iris-code comparisons.



(a) $HD = 0.45117$ (b) $HD = 0.46152$

(c) $HD = 0.47461$ (d) $HD = 0.49004$

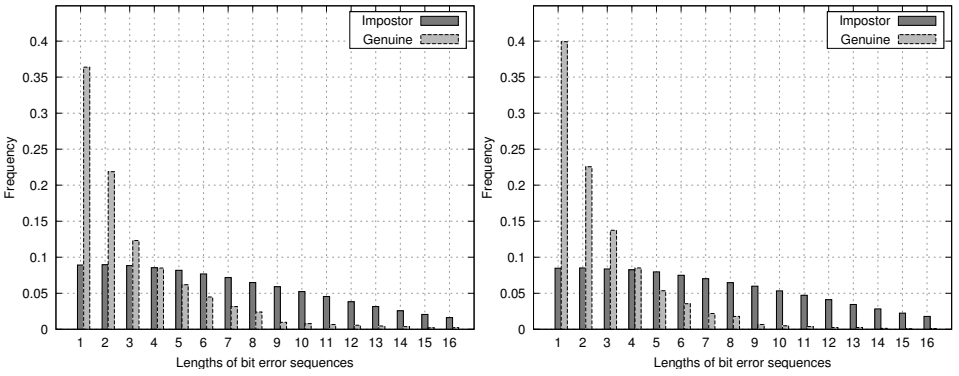
Fig. 4: Examples of bit-error patterns produced by four impostor iris-code comparisons.

$$HD = \frac{\|(codeA \oplus codeB) \cap maskA \cap maskB\|}{\|maskA \cap maskB\|}. \quad (1)$$

Experiments are conducted on the CASIAv4-Interval [CA17] and the BioSecure [Or10] iris database. Example images of both datasets are depicted in Fig. 1. An overview of the used training sets (left eye images) and testing sets (right eye images) is shown in Table 1. In experiments training and testing will be performed within and across both used databases.

3.2 Iris-Code Bit-Error Pattern Analysis

It is well known that bits in iris-codes are not mutually independent [Da04]. This is due to the internal spatial correlations within iris textures and the nature of employed filters [Da16]. Mis-matching bits between genuine iris-codes have been found to occur at boundaries of consecutive 0-bit or 1-bit sequences [HBF09, Da16]. That is, even for large *HD* scores lengths of sequences of consecutive mis-matching non-masked bits are expected to



(a) CASIAv4-Interval (b) BioSecure

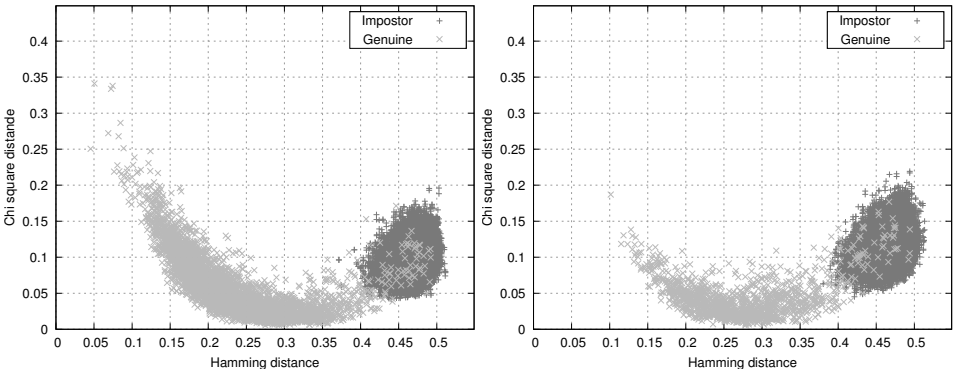
Fig. 5: Bit-error sequence lengths of bit-error patterns obtained from training sets.

be low. In contrast, for impostor comparisons these lengths tend to be higher. This is due to the facts that iris-codes of different eyes are uncorrelated and adjacent bits in iris-codes exhibit high correlation. Hence, the neighbouring bits of each non-matching bit have a high probability of being non-matching, too.

In our experiments left eye images of each database are processed in the training stage. Based on the training sets we perform all possible genuine comparisons and impostor comparisons based on the first image of each eye. Examples of bit-error patterns obtained by genuine and impostor comparisons are depicted in Fig. 3 and Fig. 4 (green pixels indicate matching bits; red pixels indicate non-matching bits). The lengths of horizontal sequences of consecutive mis-matching non-masked bits of genuine and impostor comparisons are counted and stored in separate histograms. For the training sets of the used datasets the obtained histograms are shown in Fig. 5. We observe that the frequency distributions for genuine and impostor comparisons are similar for both databases. Focusing on impostor distributions, in Fig. 5 it can be seen that, sequences of up to five consecutive mis-matching bits are almost equiprobable (also see Fig. 4). The similarity of distributions across both databases suggests that these mainly depend on the employed feature extractor (as will be shown in experimental evaluations).

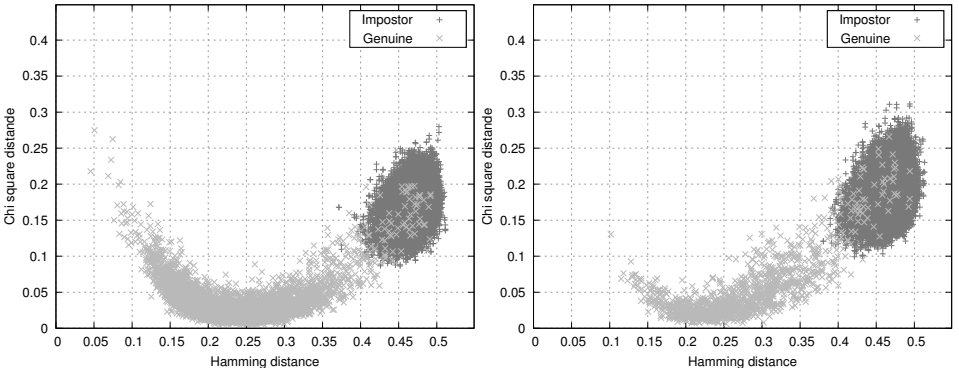
3.3 Improved Comparator

Given a pair of iris-codes, *codeA* and *codeB*, the *HD* score between them is estimated and the frequency distribution of sequences of consecutive mis-matching non-masked bits is stored in a histogram, *histAB*. This histogram is then compared against the average genuine model obtained during the training stage, *histGen*, by estimating the Chi square (χ^2) distance between both histograms,



(a) CASIAv4-Interval (b) BioSecure

Fig. 6: Scores obtained from testing sets with training performed on CASIAv4-Interval.



(a) CASIAv4-Interval (b) BioSecure

Fig. 7: Scores obtained from testing sets with training performed on BioSecure.

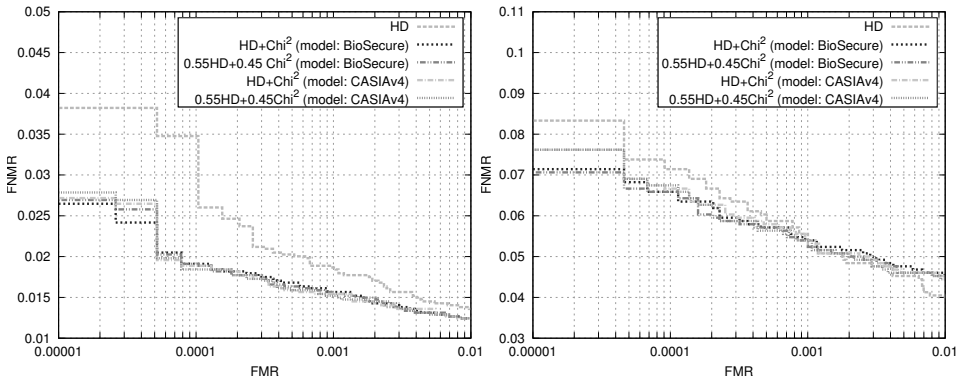
$$\chi^2(histAB, histGen) = 1/2k \sum_{i=1}^{k} (histAB_i - histGen_i)^2 / (histAB_i + histGen_i). \quad (2)$$

It has been found that the χ^2 distance is a suitable method for the proposed comparator. Alternatively, other similar methods could be employed to compare pairs of histograms, e.g. [PW10]. Note that only bit-error patterns obtained from genuine comparisons are used. No significant improvements were obtained for applying the proposed procedure to bit-error patterns produced by impostor comparisons.

Fig. 6 and Fig. 7 show scatter plots of *HD* scores and corresponding χ^2 distances for using different training sets. It can be observed that some large genuine *HD* scores still yield small χ^2 distances. Also, rather low genuine *HD* scores result in large χ^2 distance due to the small amount of bit-errors. However, as mentioned earlier, it is suggested to

Tab. 2: Performance rates (in %) obtained from the testing sets.

Comparator	Training	CASIAv4-Interval			BioSecure		
		$FNMR_{0.01}$	$FNMR_{0.001}$	$FNMR_0$	$FNMR_{0.01}$	$FNMR_{0.001}$	$FNMR_0$
HD	–	3.48	3.83	3.85	7.38	8.26	8.34
$HD+\chi^2$	CASIAv4-Interval	1.98	2.69	2.79	6.89	7.54	7.62
$0.55HD+0.45\chi^2$		1.96	2.65	2.72	6.75	7.39	7.62
$HD+\chi^2$	BioSecure	1.94	2.69	2.70	6.59	7.16	7.17
$0.55HD+0.45\chi^2$		1.92	2.63	2.65	6.56	6.99	7.14



(a) CASIAv4-Interval (b) BioSecure

Fig. 8: Detection error trade-off curves obtained from the testing sets.

estimate the χ^2 distance only for a distinct range of obtained *HD* scores, e.g. $[0.35, 0.45]$. As can be seen in Fig. 6 and Fig. 7, in such a range a diagonal line would achieve the best separation of genuine and impostor scores. That is, for a pre-defined interval the χ^2 distance is estimated as an assisting score and combined with the *HD* scores employing a weighted score-level fusion using the sum-rule. Further, we observe χ^2 distances of impostors are generally larger if the model obtained from the BioSecure training set is employed. This is because in the histogram of the BioSecure training set sequences of small lengths are weighted higher compared to the histogram of the CASIAv4-Interval database (see Fig. 5). Also, it can be seen that χ^2 distances of genuine as well as impostors are slightly larger on the BioSecure testing set. This might suggest that this database is more noisy than the CASIAv4-Interval database, which is also reflected by the obtained performance rates.

In accordance to the ISO/IEC IS 19795-1 [Int11] biometric performance is estimated in terms of false non-match rate (*FNMR*) at a targeted false match rate (*FMR*), denoted by $FNMR_{FMR}$. Obtained *FNMR*s at *FMR*s of 0.01%, 0.001% and 0% are listed in Table 2. The resulting detection error trade-off (DET) curves are shown in Fig. 8. Across considered *FMR*s the recognition accuracy is generally enhanced by the fusion of *HD* scores and χ^2 distances, which is performed within the *HD* score interval of $[0.35, 0.45]$. Due to the

fact that the histograms of bit-error sequences are similar for both databases, no significant performance drops are observed if the training is be performed on a different dataset. When using a weighted fusion only small improvements can be achieved. As an alternative to the simple (weighted) sum-rule fusion support vector machines (SVMs) could be trained to separate genuine from impostor scores.

4 Conclusions and Future Work

In this work we presented an advanced iris-biometric comparator to improve the biometric performance in an iris recognition system. In contrast to many published works, we propose an analysis of bit-error patterns produced by iris-code comparisons. In particular, we construct a model for the expected frequency distribution resulting from a genuine comparison based on a training set of iris-codes. The difference of an obtained bit-error pattern to that of the pre-trained one can be used as a second comparison score in combination with the fractional Hamming distance. At practical false match rates the recognition accuracy has be significantly improved on different databases. Reported preliminary improvements motivate further investigations of bit-error patterns of iris-code comparisons. Models of bit-errors could be, (1) constructed for different intervals of *HD* scores to improve the robustness of the proposed comparator, (2) extended to also analyse vertical, i.e. radial, correlations of bit-errors, (3) constructed for different regions of iris textures, since entropy has been found to vary significantly across iris texture regions.

Building a model for genuine bit-error patterns might be of interest for other research fields. In particular, models of bit-error patterns produced by iris-code pairs could be employed in presentation attack detection techniques [GGB16]. Moreover, machine learning techniques, e.g. convolutional neuronal networks, could be used to reliably identify error patterns produced by genuine iris-code comparisons.

Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within the Center for Research in Security and Privacy (CRISP).

References

[CA17] CASIA: , Chinese Academy of Sciences' Institute of Automation – Iris Image Database V4.0 – Interval. `http://biometrics.idealtest.org`, 2017.

[Da04] Daugman, J.: How iris recognition works. Trans. on Circuits and Systems for Video Technology, 14(1):21–30, 2004.

[Da16] Daugman, J.: Information Theory and the IrisCode. Trans. on Information Forensics and Security, 11(2):400–409, Feb 2016.

[DST11] Dong, W.; Sun, Z.; Tan, T.: Iris Matching Based on Personalized Weight Map. IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(9):1744–1757, 2011.

[GGB16] Galbally, J.; Gomez-Barrero, M.: A review of iris anti-spoofing. In: Proc. Int'l Workshop on Biometrics and Forensics (IWBF'16). pp. 1–6, 2016.

[HBF09] Hollingsworth, K. P.; Bowyer, K. W.; Flynn, P. J.: The Best Bits in an Iris Code. IEEE Trans. on Pattern Analysis and Machine Intelligence, 31(6):964–973, 2009.

[HBF11] Hollingsworth, K. P.; Bowyer, K. W.; Flynn, P. J.: Improved Iris Recognition through Fusion of Hamming Distance and Fragile Bit Distance. IEEE Trans. on Pattern Analysis and Machine Intelligence, 33(12):2465–2476, Dec 2011.

[HSH17] Hu, Y.; Sirlantzis, K.; Howells, G.: A novel iris weight map method for less constrained iris recognition based on bit stability and discriminability. Image and Vision Computing, 58:168 – 180, 2017.

[Int11] International Organization for Standardization. ISO/IEC 19795-1:2006. Information Technology - Biometric performance testing and reporting – Part 1: Principles and framework, 2011.

[Or10] Ortega-Garcia, J.; Fierrez, J.; Alonso-Fernandez, F.; Galbally, J.; Freire, M. R. et al.: The Multiscenario Multienvironment BioSecure Multimodal Database (BMDB). IEEE Trans. on Pattern Analysis and Machine Intelligence, 32(6):1097–1111, 2010.

[Pr15] Proença, H.: Iris Recognition: What Is Beyond Bit Fragility? IEEE Trans. on Information Forensics and Security, 10(2):321–332, 2015.

[PW10] Pele, O.; Werman, M.: The Quadratic-Chi Histogram Distance Family. In: 11th European Conf. on Computer Vision (ECCV'10). pp. 749–762, 2010.

[Ra16] Rathgeb, C.; Hofbauer, H.; Uhl, A.; Busch, C.: TripleA: Accelerated Accuracy-preserving Alignment for Iris-Codes. In: Proc. of the 9th IAPR/IEEE Int'l Conf. on Biometrics (ICB'16). pp. 1–8, 2016.

[RUW10] Rathgeb, C.; Uhl, A.; Wild, P.: Incremental Iris Recognition: A Single-algorithm Serial Fusion Strategy to Optimize Time Complexity. In: Proc. of the 4th IEEE Int'l Conf. on Biometrics: Theory, Application, and Systems 2010 (IEEE BTAS'10). pp. 1–6, 2010.

[RUW11] Rathgeb, C.; Uhl, A.; Wild, P.: Shifting Score Fusion: On Exploiting Shifting Variation in Iris Recognition. In: Proc. of the 26th ACM Symposium on Applied Computing (SAC'11). pp. 1–5, 2011.

[RUW12] Rathgeb, C.; Uhl, A.; Wild, P.: Iris-Biometric Comparators: Exploiting Comparison Scores towards an Optimal Alignment under Gaussian Assumption. In: Proc. of the 5th IAPR/IEEE Int'l Conf. on Biometrics (ICB'12). pp. 1–6, 2012.

[RUW13] Rathgeb, C.; Uhl, A.; Wild, P.: Iris Recognition: From Segmentation to Template Security, volume 59 of Advances in Information Security. Springer Verlag, 2013.

[US17] USIT: , University of Salzburg Iris Toolkit. `http://www.wavelab.at/sources/Rathgeb16a` Version 2.0.x, 2017.

[ZD08] Ziauddin, S.; Dailey, M. N.: Iris recognition performance enhancement using weighted majority voting. In: 15th Int'l Conference on Image Processing (ICIP'08). pp. 277–280, 2008.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 51

Domain Adaptation for CNN Based Iris Segmentation

Ehsaneddin Jalilian[1], Andreas Uhl[2], Roland Kwitt[3]

Abstract: Convolutional Neural Networks (CNNs) have shown great success in solving key artificial vision challenges such as image segmentation. Training these networks, however, normally requires plenty of labeled data, while data labeling is an expensive and time-consuming task, due to the significant human effort involved. In this paper we propose two pixel-level domain adaptation methods, introducing a training model for CNN based iris segmentation. Based on our experiments, the proposed methods can effectively transfer the domains of source databases to those of the targets, producing new adapted databases. The adapted databases then are used to train CNNs for segmentation of iris texture in the target databases, eliminating the need for the target labeled data. We also indicate that training a specific CNN for a new iris segmentation task, maintaining optimal segmentation scores, is possible using a very low number of training samples.

Keywords: Domain adaptation, CNN based iris segmentation, Iris segmentation

1 Introduction

In recent years, considerable effort has been made towards developing accurate automatic segmentation systems for variety of applications, using supervised machine learning algorithms. Accurate segmentation of iris texture in eye images is a key challenge in iris recognition, and plays vital role in accuracy of subsequent feature extraction and recognition algorithms. Application of convolutional neural networks for iris segmentation has recently received some research attention, and a few CNN based models got proposed [JU17] [Li16]. Nonetheless, as any other supervised learning model, performance of these models are highly dependent on availability of sufficient amount of labeled data. Data labeling, however, is extremely expensive and time-consuming process, especially when segmenting iris data, due to the considerable human effort involved. As a result, manually annotating large number of data for each new segmentation task (i.e. new datasets or sensors, respectively) is not a feasible choice.

In this work, we propose two domain adaptation methods to transfer the domains of source iris databases (for which segmentation labels are available) to those of the targets, generating adapted iris databases, which in turn, enable training of a Fully Convolutional Neural Network (FCN) for segmentation of iris in the target databases. Doing so, we can train a FCN for a new iris segmentation task, using adapted source iris images and their corresponding ground-truth masks, eliminating the need for the target iris ground-truth masks,

[1] Department of Computer Science, University of Salzburg, Jakob-Haringer-Str.2, Salzburg, Austria, ehsaneddin.jalilian@sbg.ac.at
[2] Department of Computer Science, University of Salzburg, Jakob-Haringer-Str.2, Salzburg, Austria, andreas.Uhl@sbg.ac.at
[3] Department of Computer Science, University of Salzburg, Jakob-Haringer-Str.2, Salzburg, Austria, roland.kwitt@sbg.ac.at1

which are extremely expensive to generate. To address this objective, we selected three publicly available iris databases, and explored their tonal distribution in terms of the intensity values at pixel-level. Subsequently, we developed a linear and also a non-linear domain adaptation hypotheses to adapt the intensity information of source databases to those of the targets, generating a set of adapted databases. Eventually, we trained a FCN with the adapted databases, and then tested it on the target databases. At the end, we evaluated the expediency of our model by comparing the segmentation results obtained by the adaptation models against those of the cross- and within-databases.

2 Related Work

Domain adaptation in computer vision is significantly focused on visual classification, with much research dedicated to generalizing across the domain transformation between images of objects and the same objects' photos in the real world [Sa10][KSD11]. In this context, many of the researches concentrated on exploring feature representations which permute the greatest distractions between two domains [Tz15][Ga16][Lo15]. Some other works tried to readjust such features by minimizing the distinction between their distributions [Lo15][Lo16]. Liu et al. proposed a coupled generative adversarial network, to learn the joint distribution of images from both the source and the target databases [LT16].

Very limited research has been conducted on domain adaptation in other key computer vision fields such as detection and segmentation. To be more precise, in detection, Hoffman et al. introduced a domain adaptation model by explicitly modeling the representation shift between classification and detection models [Ho14]. Also, in a follow-up work, they incorporated per-category adaptation using multiple instance learning [Ho15]. The detection models were later converted into FCNs for evaluating semantic segmentation performance [Ho16a]. But this work did not propose any segmentation-specific adaptation approach. The only work with focus on CNN based segmentation is proposed by Hoffman et al. [Ho16b]. They used both source and target data in a fully-convolutional domain adversarial training, minimizing the global distance of feature space between two domains. Then category updates were performed on the target images, using a constrained pixel-wise multiple instance learning objective. They used their model for semantic segmentation in city images obtained under different scenarios. The main drawback of their method is using adversarial training and shared weights. While applying this method lets the target network to adapt to the weights well, yet it degrades this process in the source network. As their experiments also show, while in most classes, they slightly improved the segmentation results, in some other classes such as "pole" and "truck" segmentation results show degradation.

3 Domain Adaptation for CNN Training

In this section, we describe our domain adaptation model for CNN based iris segmentation. Although without loss of generality, our approach is applicable to other segmentation models also. Given the source iris database X_s, and its corresponding ground-truths Y_s,

$P(X_s)$ refers to the distribution of intensities in the source iris images. Likewise, we have the target iris images X_t, and their corresponding ground-truths Y_t, while $P(X_t)$ specifies the distribution of intensities in the target iris images. Under the domain difference scenario, we assume that the conditional distributions of Y_s and Y_t are the same, but the marginal distributions of X_s and X_t differ in the two domains. The resulting distinction between the distributions in two domains is refereed to as sample bias ϕ where

$$P_t = P_s(\phi(X_s), Y_s). \tag{1}$$

Using empirical risk minimization framework for supervised learning, we want to select an optimal parameter $\psi^{'}$, to minimize the following objective function

$$\psi_t^{'} = \arg\min_{\psi \in \Psi} \sum_{(x,y) \in X \times Y} \widetilde{P}_s(\phi(X_s), Y_s) g(x, y, \psi) = \arg\min_{\psi \in \Psi} \sum_{i=1}^{N} g(\phi(x_s), y_s, \psi) \ , \tag{2}$$

where $g(x, y, \psi)$ is the loss function, and $\widetilde{P}_s(X, Y)$ is the empirical distribution of $P_s(X, Y)$. As it can be interpreted from (2), weighting the images' intensities of source data by ϕ provides the solution to the minimization function. The straight forward solution to weight the intensity values of source data is using a linear normalization model as follows:

$$b = (max(B) - min(B)) \frac{a - min(A)}{max(A) - min(A)} + (min(B)) \ , \tag{3}$$

where a and b are the input and output respectively, and $B = \{b_1, b_2, ..b_n\}$, and $A = \{a_1, a_2, ..a_n\}$. Our first (linear) domain adaptation method is based on the same model. In this way, we extracted the average range (maximum and minimum) of intensities in the iris, non-iris, and pupil regions of eye images in the target databases. Then using the above model (3), we weighted the intensity information of source databases to those of the targets, to generate new adapted databases as we already mentioned.

As it can be seen, this model provides a linear solution to our domain adaptation problem. In practice, in this method for each region, all the source intensity ranges get normalized to "a single average intensity range of that region in the target database." Yet it is a fact that, the intensity ranges of the target regions follow a non-linear distribution in the target databases. To address this non-linearity, we propose our second (non-linear) domain adaptation method. For this purpose, after extracting the maximums and minimums of each region in the target databases, we developed a probability distribution function (PDF) for each. To transfer the intensities in the source regions to those of the targets, initially we drew a random value from the corresponding PDFs, following a normal distribution.

However, this strategy seemed not to be so promising, as it neglected the complimentary relation between maximums and minimums in each region. Further analysis of the extracted intensities also revealed that there exists an obvious mutual relation between maximum and minimum intensity values in each region. So that, as the maximums increase, minimums also increase, and vise versa. To address this relation, after extracting the intensity ranges, for each unique maximum value, we calculated the mean of corresponding minimum values. Then we developed a cross-value matrix for each region, using

these two variables. Next, we applied kernel smoothing regression to this data to generate a polynomial function $f(X)$ as follows:

$$f(x) = p_1 x^n + p_2 x^{n-1} + ... + p_n x + p_{n+1} \ , \tag{4}$$

where x represents the input (minimum) to the model, and n is the degree of polynomial function. Now, to adapt the domain of each source image to that of the targets, we randomly selected a minimum for each region, and then estimated its corresponding maximum using the polynomial model we proposed, as demonstrated in figure 1. As a result, unlike in the linear adaptation method, where all images were mapped to the same range, here each adapted image has a potentially different range.

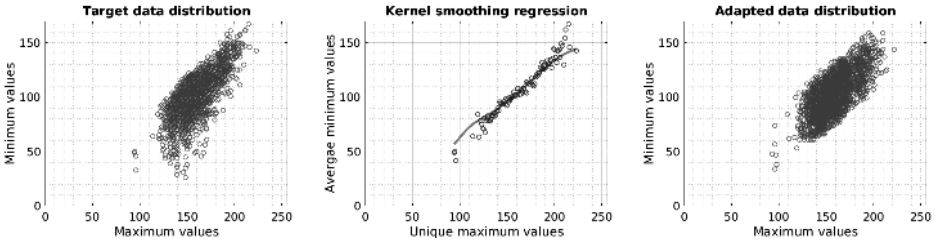


Fig. 1: Sample non-linear data adaptation steps

4 Experimental Framework

To assess the expediency of our domain adaptation methods we carried out a set of segmentation experiments on the databases. The details of these experiments are explained in the next section respectively. Yet, in the following we explain the framework for these experiments.

4.1 Databases

For our segmentation experiments we used three publicly available iris databases. The Casia-iris-interval-v4 (Casia4i) database [4], which contains 2640 iris images belonging to 249 subjects. The iris images in this database were acquired under near-infrared illumination. The IITD database (Iitd) [5], which consists 2240 iris images corresponding to 224 subjects. All these images are acquired in indoor environment, in near-infrared illumination. The Casia-iris-aging-v5 database, which is a subset of the upcoming Casia-v5 (Casia5a) iris database [6], contains 120 images per eye and user from video sequences captured in 2009, and 20 images per eye and user from video sequences captured in 2013. The segmentation ground-truth masks for these databases were provided by the University of Salzburg [7].

[4] Chinese Academy of Sciences, Institute of Automation, Center for biometrics and security research, http://biometrics.idealtest.org

[5] Indian Institute of Technology Delhi, IIT Delhi Iris Database, http://www4.comp.polyu.edu.hk/ csajaykr/-database.php

[6] see http://www.biometrics.idealtest.org

[7] http://www.wavelab.at/sources/Hofbauer14b

4.2 FCN Netwrok

The architecture of the network we used in this work is similar to the basic fully convolutional encoder-decoder network proposed by Kendall et al. [BKC15]. However, we redesigned the softmax layer to segment the iris and non-iris areas only. The network's encoder architecture is organized in four stocks, containing a set of blocks. Each block comprises a convolutional layer, a batch normalization layer, and a rectified-linear non-linearity layer. The corresponding decoder architecture, likewise, is organized in four stocks of blocks, whose layers are similar to those of the encoder blocks, except that here each block includes an up-sampling layer also. The decoder network ends up to a softmax layer which generates the final segmentation map. More details about the technical specification of the network and layers can be found in the relevant reference. The network was implemented in the "Caffe" deep learning framework.

4.3 Metrics and Measurements

We estimated iris segmentation accuracies using two segmentation error scores of nice1 $(n1)$ and nice2 $(n2)$, which are based on the NICE.I protocol [8]. The error score nice1 calculates the proportion of corresponding disagreeing pixels (by the logical exclusive-or operator) over all the image as follows:

$$nice1 = \frac{1}{c \times r} \sum_{c'} \sum_{r'} O(c', r') \otimes C(c', r') \,, \tag{5}$$

where c and r are the columns and rows of the segmentation masks, and $O(c', r')$ and $C(c', r')$ are, respectively, pixels of the output and the ground-truth mask. The error score nice2 intends to compensate the disproportion between the priori probabilities of iris and non-iris pixels in the images - it averages type-I and type-II errors, i.e. between the fp (false positives) and fn (false negatives) rates as follows:

$$nice2 = \frac{1}{2} \left(fp + fn \right) . \tag{6}$$

Additionally, we considered the F score $(f1)$ to estimate iris segmentation accuracies also. The F score is the harmonic mean of precision (P) (the fraction of relevant instances among the retrieved instances) and recall (R) (the fraction of relevant instances that have been retrieved over total relevant instances) as follows:

$$f1 = 2 \, \frac{RP}{R + P} \,. \tag{7}$$

[8] http://nice1.di.ubi.pt/dates.htm

The values of nice1 and nice2 are bounded in [0, 1] interval, and ”1” and ”0” are respectively the worst and the best scores. The F score values are bounded in [1, 0] interval, and ”0” and ”1” are the worst, and the best scores respectively.

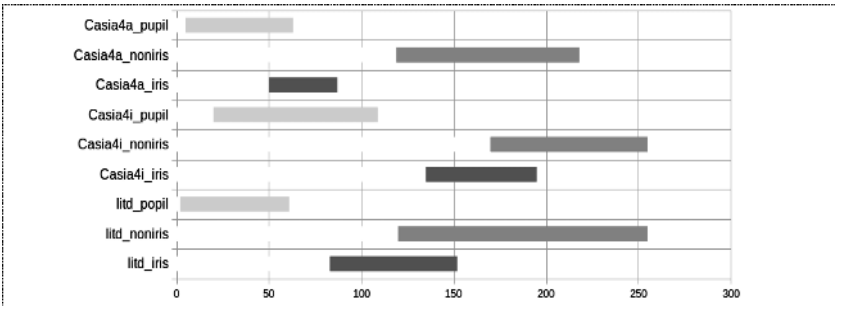


Fig. 2: Average intensity ranges of iris, non-iris, and pupil regions in databases

5 Experiments and Discussions

We evaluated the eminence of our domain adaptation model by a set of experiments. In this way, initially we developed six sets of unique database pairs (source-target), using three available databases. Next, we explored the distributions of domains in the target databases, extracting the intensity ranges of iris, non-iris, and pupil regions of eye images in these databases (figure 2 reflects these information). Then, using our domain adaptation methods, we transfered the intensity values of the specified regions in source databases to those of the targets, to produce an adapted database for each pair. Next, we trained our network with each adapted database, and then tested it on the corresponding target databases (adapted-target). Figure 3 and figure 4 show sample adapted images and their corresponding segmentations results for three database pairs, applying linear and non-linear adaptation methods respectively.

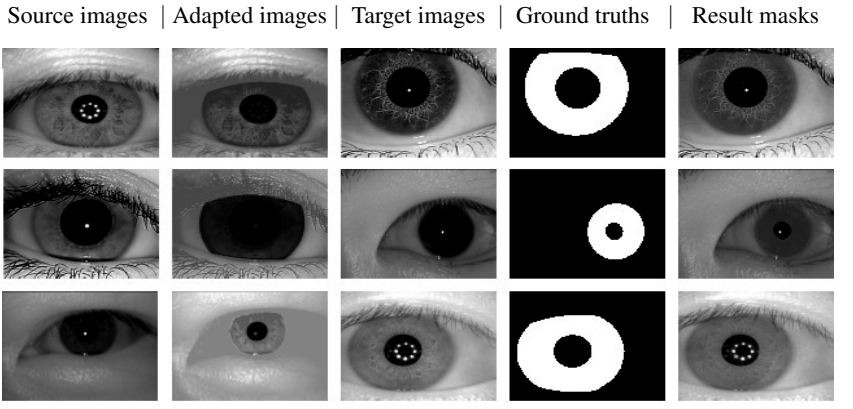


Fig. 3: Sample adapted images and their corresponding segmentation results for Casia4i-Iitd (first row), Iitd-Casia5a (second row), and Casia5a-Casia4i (third row) database pairs (source-target) using the linear domain adaptation method

The segmentation results then were compared against the baseline results (results of applying network trained with the source databases directly to the target databases without adaptation). Table 1 shows the segmentation scores for our linear-based (LB), and non-linear-based (NB) adaptation methods against the baseline (source-target) results.

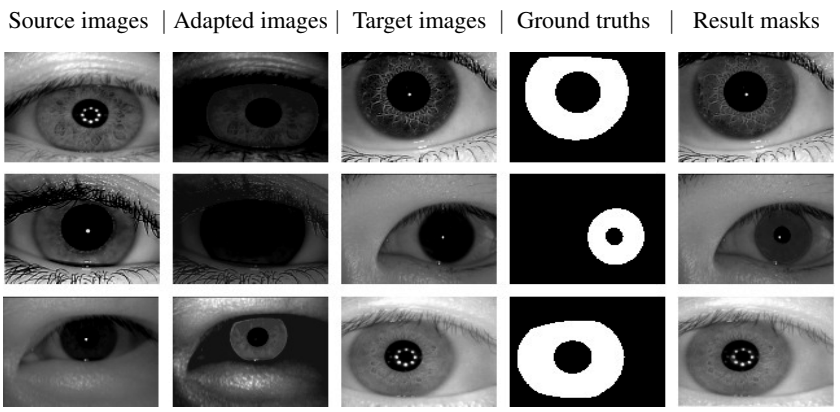


Fig. 4: Sample adapted images and their corresponding segmentation results for Casia4i-Iitd (first row), Iitd-Casia5a (second row), and Casia5a-Casia4i (third row) database pairs (source-target) using the non-linear domain adaptation method

In addition, figure 5 provides further information including: min, max, median, quantiles, and outliers for the liner-based adaptation experiments in form of box-plots

Method	Adapted-target (LB)			Adapted-target (NB)			Baseline(Source-target)		
Scores	nice1	nice2	f1	nice1	nice2	f1	nice1	nice2	f1
Casia5a-casia4i	0.186	0.220	0.610	0.274	0.353	0.098	0.292	0.640	0.003
Casia5a-iitd	0.148	0.172	0.781	0.266	0.305	0.498	0.229	0.221	0.473
Casia4i-casia5a	0.066	0.194	0.730	0.027	0.074	0.859	0.274	0.406	0.341
Casia4i-iitd	0.121	0.141	0.808	0.102	0.095	0.812	0.218	0.219	0.724
Iitd-casia5a	0.062	0.185	0.739	0.034	0.088	0.813	0.049	0.117	0.830
Iitd-casia4i	0.299	0.319	0.569	0.208	0.174	0.374	0.315	0.584	0.045

Tab. 1: Segmentation scores for the linear-based (LB), and non-linear-based (NB) domain adaptation methods against the baseline (source-target) results

As the experiment results in Table 1 show, almost all linear domain adaptations result in significant improvement of iris segmentations compared to the baseline results. Slightly lower, yet stable improvements can also be seen in the segmentation results of non-linear domain adaptations. It should be noted that feature representations affecting the weights during training process are not limited to tonal distributions, and further features such as geometric properties of iris, non-iris, and pupil regions are definitely affecting this process. Here we just considered the tonal distributions, so the results are not comparable with the optimal solution when directly training with the target dataset. All in all, the overall results confirm the key conclusion that tonal distribution (intensity ranges of iris, non-iris, and pupil) plays a key role in generalization of FCNs on new iris data that differs from the training data. It is also interesting to note that, while the segmentation results for

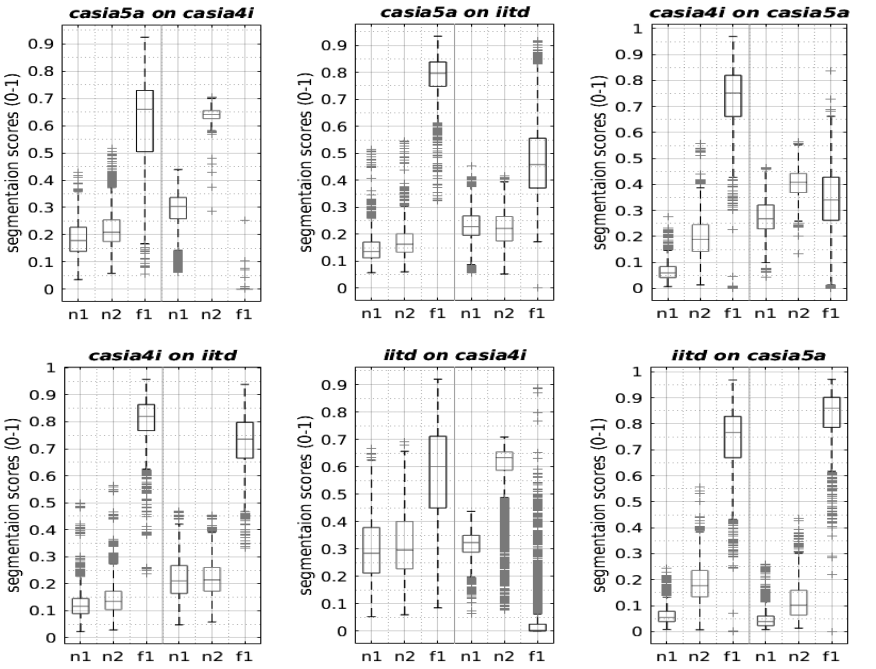


Fig. 5: Segmentation results for the linear domain adaptations (left side of graphs), against the baseline results (right side of graphs)

linearly adapted Iitd-casia5a databases show slightly lower scores than the baseline, yet the segmentation results for non-linear adaptation of the same databases score much better compared to those of the baseline. Similar affinity can be found in the segmentation results of Casia5a-iitd databases, but in reverse manner.

Method	Target-Target		
Scores	nice1	nice2	f1
Casia5a-casia5a	0.019	0.038	0.925
Casia4i-casia4i	0.033	0.038	0.937
Iiitd-iitd	0.027	0.032	0.951

Tab. 2: Optimal (target-target) segmentation results

While the proposed domain adaptation methods proved to effectively transfer the domains between the iris databases, yet the segmentation results obtained are far from the optimal iris segmentation scores as demonstrated in Table 2. To this extent, with the aim of minimizing the number of labeled data required to train the CNNs for new iris segmentation tasks, and maintaining optimal segmentation scores, we conducted a series of additional experiments. In this way, we decreased the number of labeled samples required to train a CNN for a new iris segmentation task stepwise, obeying the framework we used for our optimal (target-target) experiments. Table 3 demonstrates the results for these experiments.

Considering the optimal segmentation results in Table 2, we can see that for most databases optimal segmentation scores can be achieved using maximum number of 100 training samples. However, in most cases slightly lower, but very close scores can be achieved with 50 or even 25 samples.

Database	Casia5a			Casia4i			Iiitd		
Score	nice1	nice2	f1	nice1	nice2	f1	nice1	nice2	f1
15 pcs	0.075	0.082	0.875	0.205	0.263	0.502	0.089	0.097	0.856
25 pcs	0.064	0.077	0.896	0.099	0.115	0.814	0.077	0.083	0.879
50 pcs	0.050	0.070	0.909	0.078	0.068	0.841	0.063	0.070	0.889
100 pcs	0.021	0.040	0.921	0.038	0.039	0.926	0.035	0.037	0.941

Tab. 3: Segmentation results for decreased number of training samples

6 Conclusion

Application of convolutional neural networks for iris segmentation has recently received first research attention, and some CNN based models got introduced for this purpose by researchers. Nonetheless, as any other supervised learning model, training these models require adequate amount of labeled iris data. Due to the significant human effort involved, preparing labeled data to train these networks for new segmentation tasks is very expensive and time consuming. In this work, we proposed two adaptation methods to transfer the domains of source iris databases to those of the targets, producing adapted databases. The adapted iris images along with their corresponding ground-truth masks then enabled training of a FCN network for segmentation in target iris databases, eliminating the need for the target ground-truth masks.

While experimental results proved expediency of these two methods, yet in some cases, their segmentation scores were far from the optimals. With the aim of minimizing the number of labeled iris images required to train the network for new iris segmentation tasks, and also maintaining optimal segmentation scores, we decreased the number of training samples stepwise as an alternative approach to domain adaptation. The experiments demonstrated that for most databases, optimal segmentation scores can be achieved using maximum of 100 training data samples. In our future work, we will investigate the relations between the two proposed adaptation methods and the reasons for the different results. Beside this, we will explore more feature representations which encourage maximal distinction between two domains, hoping to be able to develop a more comprehensive domain adaptation method.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 700259.

References

[BKC15] Badrinarayanan, Vijay; Kendall, Alex; Cipolla, Roberto: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561, 2015.

[Ga16] Ganin, Yaroslav; Ustinova, Evgeniya; Ajakan, Hana; Germain, Pascal; Larochelle, Hugo; Laviolette, François; Marchand, Mario; Lempitsky, Victor: Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59):1–35, 2016.

[Ho14] Hoffman, Judy; Guadarrama, Sergio; Tzeng, Eric S; Hu, Ronghang; Donahue, Jeff; Girshick, Ross; Darrell, Trevor; Saenko, Kate: LSDA: Large Scale Detection through Adaptation. In (Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; Weinberger, K. Q., eds): Advances in Neural Information Processing Systems 27, pp. 3536–3544. Curran Associates, Inc., 2014.

[Ho15] Hoffman, Judy; Pathak, Deepak; Darrell, Trevor; Saenko, Kate: Detector discovery in the wild: Joint multiple instance and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2883–2891, 2015.

[Ho16a] Hoffman, Judy; Pathak, Deepak; Tzeng, Eric; Long, Jonathan; Guadarrama, Sergio; Darrell, Trevor; Saenko, Kate: Large scale visual recognition through adaptation using joint representation and multiple instance learning. Journal of Machine Learning Research, 17(142):1–31, 2016.

[Ho16b] Hoffman, Judy; Wang, Dequan; Yu, Fisher; Darrell, Trevor: FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation. arXiv preprint arXiv:1612.02649, 2016.

[JU17] Jalilian, Ehsaneddin; Uhl, Andreas: Iris Segmentation Using Fully Convolutional Encoder–Decoder Networks. In: Deep Learning for Biometrics, pp. 133–155. Springer, 2017.

[KSD11] Kulis, Brian; Saenko, Kate; Darrell, Trevor: What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, pp. 1785–1792, 2011.

[Li16] Liu, Nianfeng; Li, Haiqing; Zhang, Man; Liu, Jing; Sun, Zhenan; Tan, Tieniu: Accurate iris segmentation in non-cooperative environments using fully convolutional networks. In: Biometrics (ICB), 2016 International Conference on. IEEE, pp. 1–8, 2016.

[Lo15] Long, Mingsheng; Cao, Yue; Wang, Jianmin; Jordan, Michael: Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning. pp. 97–105, 2015.

[Lo16] Long, Mingsheng; Zhu, Han; Wang, Jianmin; Jordan, Michael I: Unsupervised domain adaptation with residual transfer networks. In: Advances in Neural Information Processing Systems. pp. 136–144, 2016.

[LT16] Liu, Ming-Yu; Tuzel, Oncel: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 469–477, 2016.

[Sa10] Saenko, Kate; Kulis, Brian; Fritz, Mario; Darrell, Trevor: Adapting visual category models to new domains. Computer Vision–ECCV 2010, pp. 213–226, 2010.

[Tz15] Tzeng, Eric; Hoffman, Judy; Darrell, Trevor; Saenko, Kate: Simultaneous deep transfer across domains and tasks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4068–4076, 2015.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 61

SIC-Gen: A Synthetic Iris-Code Generator

Pawel Drozdowski[1,2], Christian Rathgeb[2], Christoph Busch[2]

Abstract: Nowadays large-scale identity management systems enrol more than one billion data subjects. In order to limit transaction times, biometric indexing is a suitable method to reduce the search space in biometric identifications. Effective testing of such biometric identification systems and biometric indexing approaches requires large datasets of biometric data. Currently, the size of the publicly available iris datasets is insufficient, especially for system scalability assessments. Synthetic data generation offers a potential solution to this issue; however, it is challenging to generate data that is both statistically sound and visually realistic - for the iris, the currently available approaches prove unsatisfactory.

In this paper, we present a method for generation of synthetic binary iris-based templates, i.e. Iris-Codes, which are the *de facto* standard used throughout major biometric deployments around the world. We validate the statistical properties of the synthetic templates and show that they closely resemble ones produced from real ocular images. With the proposed approach, large databases of synthetic Iris-Codes with flexibly adjustable properties can be generated.

Keywords: Biometrics, Iris Recognition, Iris-Code, Synthetisation

1 Introduction

The iris is one of the most widely applied biometric modalities. In recent years, several large-scale deployments have been created, most notably the Indian National ID program [Un10], which has, at the time of this writing, enrolled over one billion subjects with biometric data including the irides. Despite using efficient comparators (e.g. Hamming distance for the iris) and parallelism, the computational load faced by such deployments in the identification scenario is extremely high. With biometric workload reduction as a motivation, many approaches for indexing of iris data have been developed [PN17]. However, evaluation of such approaches and their scalability is often questionable due to lack of large test datasets. While various publicly available iris databases with near-infrared (NIR) data exist, they are relatively small. At the time of this writing, some of the largest publicly available datasets, CASIA-IrisV4-Thousand and ND-CrossSensor-Iris-2013, contain merely 20.000 images from 1000 subjects and 146.550 images from 676 subjects, respectively. This is several orders of magnitude smaller than some of the large-scale deployments nowadays.

Synthetic data generation is one possible way of dealing with the issue of testing efficient indexing methods. Most of the existing approaches for synthetic iris generation attempt to synthesise an entire iris image or texture [Le03, Cu04, MR05, WSG05, ZS05, SR06,

[1] Norwegian Biometrics Laboratory, NTNU, Gjøvik, Norway
[2] da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany {pawel.drozdowski,christian.rathgeb,christoph.busch}@h-da.de

ZSC07, WTS08, PN13]. The main issues with such approaches include the computational costs and the difficulty in guaranteeing the statistical properties of the real data. The vast majority of operational iris biometric systems are based on the Iris-Code [Da04], making it a *de facto* standard. Generating Iris-Codes (feature vectors) directly is therefore also viable and may offer better control over the statistical properties of the synthetic data. Recently, two such approaches have been proposed. Proença and Neves [PN13] provide a method of Iris-Code synthesis based on bit correlations; the method is shown to attain some of the desired statistical properties (the shapes of the genuine and impostor distributions). It is also somewhat flexible with adjustable parameters; however, it does not allow to generate a set of templates following a desired score distribution. Furthermore, the filter response resulting from the typical feature extraction process is not modelled (in other words, the produced synthetic Iris-Codes scantily resemble the ones produced from real iris images through the commonly used iris processing pipeline). Lastly, typical error patterns between two mated templates are not modelled. Daugman [Da16] proposed to use a simple hidden Markov model to generate a stream of bits and showed that it can be adjusted, so that the produced templates mimic the impostor distribution of real iris templates. However, the produced streams are 1-dimensional (i.e. do not model the correlation between the Iris-Code rows); furthermore, the method does not offer a way to generate more than one template per subject (i.e. it is not possible to use it for simulating genuine comparisons). As such, it might only be useful for stress-testing of iris identification systems.

In this paper, we present a synthetic Iris-Code generator, which both reflects the statistical properties of the real Iris-Codes and resembles the real templates visually. An important feature of the proposed approach is its flexibility, in that it allows to generate Iris-Codes with an arbitrary resolution and an arbitrary score distribution of mated templates, unlike any of the approaches currently in the literature. To facilitate reproducible research, the software written in Python3 programming language, is released to the scientific community under a permissive license.

The remainder of this paper is organised as follows: section 2 describes the proposed method of synthetic Iris-Code generation. In section 3 the properties of the generated templates are validated, while section 4 contains concluding remarks.

2 Proposed Method

When generating synthetic Iris-Codes, several matters have to be taken into account:

- Dataset

  Score distributions The distributions of Hamming distance scores must closely resemble the ones produced by real data.

  Degrees of freedom Based on a large number of comparison scores from non-mated templates, the effective number of independent bits (degrees of freedom) can be calculated. Degrees of freedom can be seen as discrimination entropy as a measure of information content in iris images and has to be close to that of the real data.

- Individual templates

  Bit correlation The bits in an Iris-Code are far from independent. There exist correlations between both rows and columns, which result in long sequences of identical consecutive bits. The reason for this is partially the anatomy of iris patterns, as well as the nature of the commonly used feature extractors [Da16]. Those correlations have to be reflected in the synthetic data.

  Error patterns The majority of bit mismatches between two mated Iris-Codes occurs for bits resulting from wavelet response close to 0 (i.e. where the response phase changes). Those occur mostly on the edges of the bit sequences, and are called the ”fragile” bits [HBF09]. They have to be present in the synthetic data. Additional noise sources, such as the occlusions resulting from the eyelids, have to be modelled as well.

  Rotation In the real data, rotations of the eye, which are mainly caused by head tilts (i.e. roll pose), potentially result in misalignment between two mated samples. In Iris-Codes, this is represented by circular horizontal shifts of the matrix columns, which have to be modelled in the synthetic data.

The proposed generator synthesises Iris-Codes as pairs of mated templates, referred to as Iris-Codes *IC1* and *IC2* in the algorithm description and figure 1 below. The bold-filled arrows denote the changes to the template throughout the process, while the thin arrows denote the system parameters.

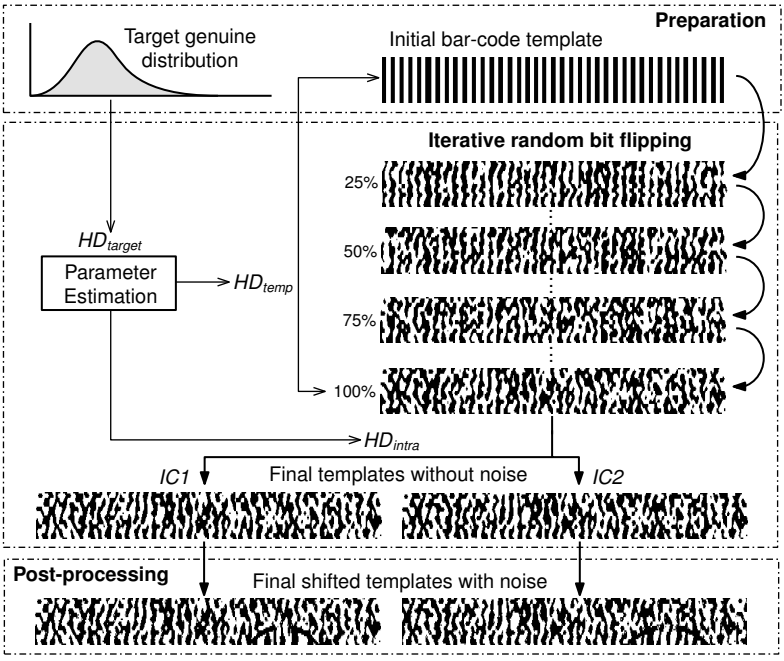


Fig. 1: The process of generating an Iris-Code pair with SIC-Gen

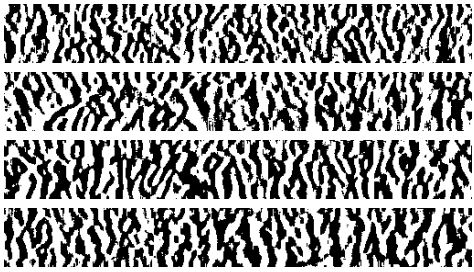
1. **Preparation**, during which a base Iris-Code matrix is created as follows:
   - The first row is created by generating alternating sequences of 0's and 1's with lengths drawn from a normal distribution. The distribution parameters can be estimated empirically, by measuring the sequence lengths in real Iris-Codes.
   - By duplicating that row, a simple bar-code pattern is generated.
2. **Parameter Estimation**, during which system configuration variables are calculated based on the user input.
   - A target Hamming distance (HD_{target}) between *IC1* and *IC2* is drawn from a random distribution.
   - HD_{temp} and HD_{intra} (see figure 1 and next step of the process description), are estimated based on HD_{target}. Following relations are satisfied: $HD_{temp} + HD_{intra} = C$ and $HD_{target} = 2HD_{intra} - O$, where O is the expected overlap of bit mismatches introduced by the process described in the next step; C remains constant for a batch of generated templates, and affects the effective number of independent bits (degrees of freedom) in the synthetic data.
3. **Iterative bit flipping**, during which a pair of mated Iris-Code templates is created from the base Iris-Code.
   - The bits at the edges of consecutive bit sequences (i.e. where sequences of 1's turn to 0's and vice versa) are randomly flipped. After HD_{temp} from the original bar-code template is reached, the template is split into *IC1* and *IC2*. Subsequently, bit flipping occurs until HD_{intra} between them is reached.
   - Additionally, majority voting and median filtering are applied to make the patterns visually smoother. Furthermore, the chances of bit flips are adjusted on per-row basis to simulate the collarette and furrow structures in real irides.
   - This step can be accelerated by applying an initial shifting pattern to the bar-code template produced in step 1.
4. **Post-processing**, during which additional noise factors are accounted for. Those include:
   - Adding the characteristic pattern resulting from an eyelid, as well as the noise beneath it.
   - Adding additional noise in the row near the pupil and simulating occlusions.
   - Storing the noise masks.
   - Applying circular shifts to the Iris-Code to simulate sample roll pose.

The process generates Iris-Codes of a default size; smaller sizes, if desired, are sampled from this size. The default dimension is motivated by the ISO/IEC international standard on Biometric sample quality [IS15]. There, the minimum iris radius is recommended to be at least 80 pixels (for the smallest reported human iris), which corresponds to a texture width of $80 * 2\pi \approx 502$ pixels when unrolled. The recommended optimal iris-pupil ratio is

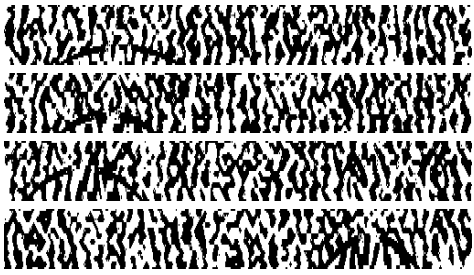
0.2, which corresponds to a pupil of $80 * 0.2 = 16$ pixels, and thus an iris texture of 64 rows. Thus, the default size of the generated Iris-Codes is 512×64 bits. There are numerous adjustable parameters, which allow to mimic different properties of the Iris-Code (e.g. the correlations between rows and columns, noise). Notably, it is also possible to *guarantee* an arbitrary distribution of genuine scores and thereby simulate sample quality. For the data generated in this paper, the HDs are drawn from a Weibull distribution, due to its close resemblance to real data; another candidate could be the Gamma distribution. Yet another approach could be to empirically estimate a distribution from real data and use it instead.

3 Validation

In this section, the properties of the synthetically generated data are validated with respect to the requirements outlined in section 2. The visual comparison between real and synthetic Iris-Codes can be seen in figure 2. The real Iris-Codes were produced by using the OSIRIS toolkit [ODGS16] to process the near-infrared images from the iris subset of the BioSecure [Or10] database. The toolkit provides the commonly used 2D-Gabor feature extraction algorithm to produce the Iris-Codes. The synthetic Iris-Codes bear an excellent resemblance to the real ones.



(a) Real

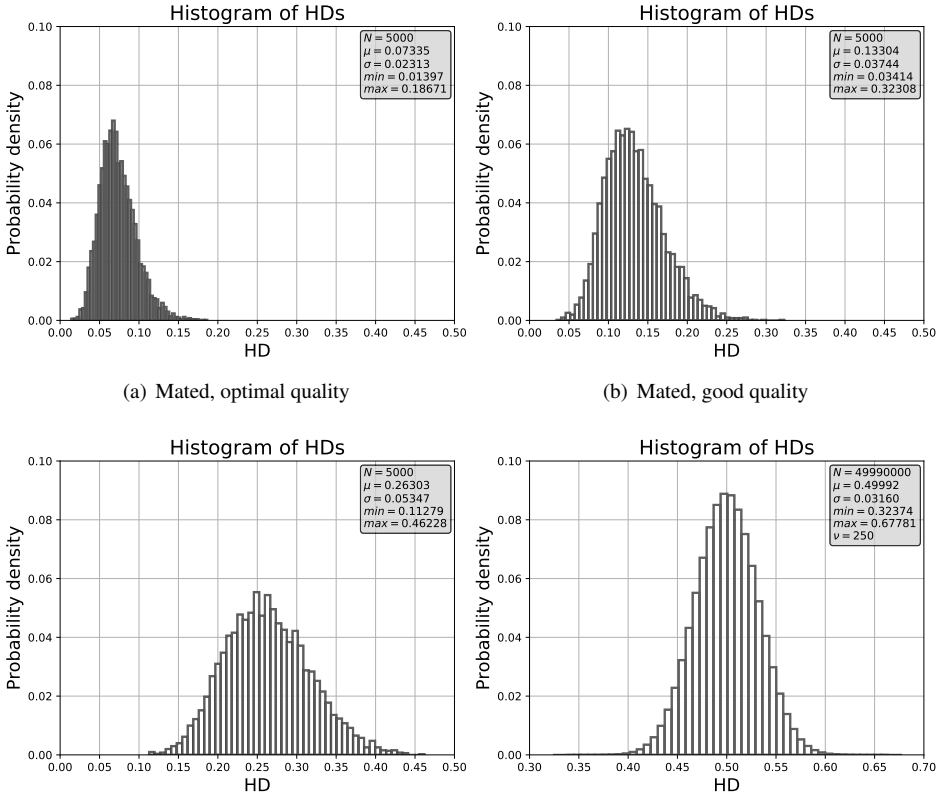


(b) Synthetic

Fig. 2: Example Iris-Codes produced from real eye images and generated by the proposed method

After confirming the visual appearance of the synthetic Iris-Codes to closely resemble that of the real data, their statistical properties are validated. Figure 3(d) shows the distribution of scores for non-mated templates for a large number of comparisons (N). The resulting distribution and its statistical properties (the yellow box in the image), including degrees of

freedom (ν), are identical to that exhibited by the real data, shown by Daugman in [Da04]. In figures 3(a), 3(b) and 3(c), example distributions of comparison scores for mated templates are shown, representing simulating of optimal, good and non-optimal quality data, respectively. As mentioned earlier, the mated distributions can be specified arbitrarily due to the nature of the template generation process (see section 2). The score distributions in figure 3 were produced using Iris-Codes of size 256×8 bits (same as used by Daugman in the paper cited above), sampled from the default size Iris-Codes generated by the process described in the previous section.



(a) Mated, optimal quality

(b) Mated, good quality

(c) Mated, non-optimal quality

(d) Non-mated

Fig. 3: Distributions of Hamming distances for a large number of comparisons between synthetic templates

Due to correlations between bits in an Iris-Code, its rows comprise of sequences of consecutive identical bits. It is of interest to verify, that the synthetic data follows that property. As real data reference, sequence lengths for all templates from the iris subset of the BioSecure database were computed. In figure 4, those distributions are shown, along with sequence lengths produced by Daugman's HMM from [Da16]. The distribution for the synthetic data generated by SIC-Gen closely follows the one exhibited by the real data.

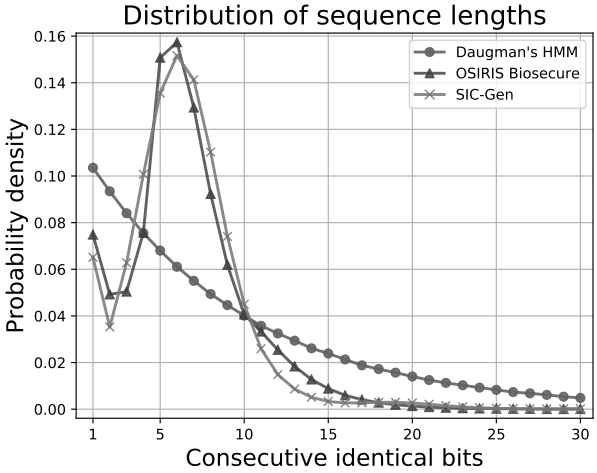
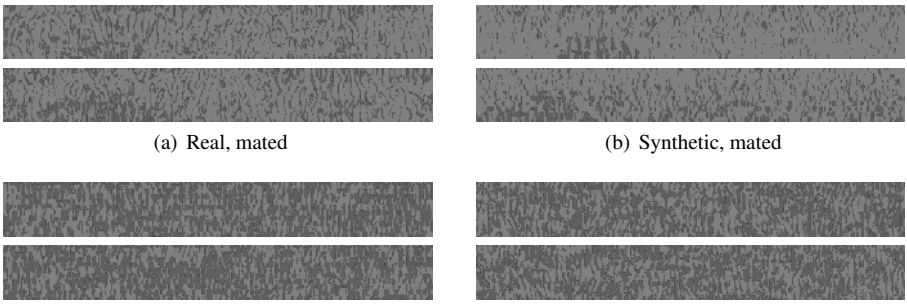


Fig. 4: Visualisation of lengths of sequences of consecutive bits in real data from BioSecure database, SIC-Gen synthetic templates and synthetic templates generated with Daugmann's HMM

Figure 5 shows example error patterns for comparisons between mated and non-mated templates. For the mated template pairs, the bit mismatches occur at the edges of sequences of consecutive identical bits, resulting in the pattern akin to that shown in real data by Hollingsworth *et al.* [HBF09].



(a) Real, mated (b) Synthetic, mated

(c) Real, non-mated (d) Synthetic, non-mated

Fig. 5: Example error patterns for comparisons between the real Iris-Codes from the BioSecure dataset and between the synthetic Iris-Codes

4 Conclusion and Future Work

In this paper, a method for generating synthetic Iris-Codes has been presented. The proposed method allows for a flexible specification of the score distribution between mated templates, to allow simulating different sample quality, acquisition environments etc.; the bit mismatches between two mated templates follow the so-called "fragile bits" patterns

observed in real data. Simultaneously, the important statistical properties (e.g. degrees of freedom) of the distribution of non-mated comparison scores are maintained. Additionally, the synthetic Iris-Codes resemble the real ones visually. They reflect the correlations between Iris-Code bits resulting in long sequences of consecutive identical bits, as well as the typical noise sources, such as the eyelid pattern, circular shifts, wavelet noise and additional noise near the pupil. By accounting for all the aforementioned statistical and visual properties of real iris data, the proposed method represents a significant improvement over the current state-of-the-art and can be used in research cases where large iris datasets are needed, but unavailable. In future work, the authors intend to employ the synthetic Iris-Codes in large-scale testing of biometric indexing approaches, as well as to attempt to generate iris textures and/or images from the synthetic data using learning-based methods, e.g. Galbally *et al.* [Ga13].

Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within Center for Research in Security and Privacy (CRISP).

References

[Cu04] Cui, J.; Wang, Y.; Huang, J.; Tan, T.; Sun, Z.: An iris image synthesis method based on PCA and super-resolution. In: 17th Intl. Conf. on Pattern Recognition. volume 4, pp. 471–474, August 2004.

[Da04] Daugman, J.: How iris recognition works. IEEE Trans. on Circuits and Systems for Video Technology, 14(1):21–30, January 2004.

[Da16] Daugman, J.: Information theory and the IrisCode. IEEE Trans. on Information Forensics and Security, 11(2):400–409, February 2016.

[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, October 2013.

[HBF09] Hollingsworth, K. P.; Bowyer, K. W.; Flynn, P. J.: The best bits in an Iris Code. IEEE Trans. on Pattern Analysis and Machine Intelligence, 31(6):964–973, June 2009.

[IS15] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 29794-6:2015. Information technology – Biometric sample quality – Part 6: Iris image data. International Organization for Standardization and International Electrotechnical Committee, July 2015.

[Le03] Lefohn, A.; Budge, B.; Shirley, P.; Caruso, R.; Reinhard, E.: An ocularist's approach to human iris synthesis. IEEE Computer Graphics and Applications, 23(6):70–75, November 2003.

[MR05] Makthal, S.; Ross, A.: Synthesis of iris images using Markov random fields. In: 13th European Signal Processing Conf. pp. 1–4, September 2005.

[ODGS16] Othman, N.; Dorizzi, B.; Garcia-Salicetti, S.: OSIRIS: An open source iris recognition software. Pattern Recognition Letters, 82(2):124–131, September 2016.

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

[PN13] Proença, H.; Neves, J. C.: Creating synthetic IrisCodes to feed biometrics experiments. In: IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications. pp. 8–12, September 2013.

[PN17] Proença, H.; Neves, J. C.: Iris biometric indexing. In (Rathgeb, C.; Busch, C., eds): Iris and periocular biometric recognition, p. 25. IET, 2017.

[SR06] Shah, S.; Ross, A.: Generating synthetic irises by feature agglomeration. In: Intl. Conf. on Image Processing. pp. 317–320, October 2006.

[Un10] Unique Identification Authority of India (UIDAI): , Aadhaar issued summary. `https://portal.uidai.gov.in/uidwebportal/dashboard.do`, 2010. Last accessed: 2017-07-31.

[WSG05] Wecker, L.; Samavati, F.; Gavrilova, M.: Iris synthesis: a reverse subdivision application. In: 3rd Intl. Conf. on computer graphics and interactive techniques in Australasia and South East Asia. pp. 121–125, November 2005.

[WTS08] Wei, Z.; Tan, T.; Sun, Z.: Synthesis of large realistic iris databases using patch-based sampling. In: 19th Intl. Conf. on Pattern Recognition. pp. 1–4, December 2008.

[ZS05] Zuo, J.; Schmid, N. A.: A model based, anatomy based method for synthesizing iris images. In: Intl. Conf. on Biometrics. pp. 428–435, January 2005.

[ZSC07] Zuo, J.; Schmid, N. A.; Chen, X.: On generation and analysis of synthetic iris images. IEEE Trans. on Information Forensics and Security, 2(1):77–90, March 2007.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 71

xTARP: Improving the Tented Arch Reference Point Detection Algorithm

Johannes Merkle[1], Benjamin Tams[2], Benjamin Dieckmann[3], Ulrike Korte[4]

Abstract: In 2013, Tams et al. proposed a method to determine directed reference points in fingerprints based on a mathematical model of typical orientation fields of tented arch type fingerprints. Although this *Tented Arch Reference Point* (TARP) method has been used successfully for pre-alignment in biometric cryptosystems, its accuracy does not yet ensure satisfactory error rates for single finger systems.

In this paper, we improve the TARP algorithm by deploying an improved orientation field computation and by integrating an additional mathematical model for arch type fingerprints. The resulting *Extended Tented Arch Reference Point* (xTARP) method combines the arch model with the tented arch model and achieves a significantly better accuracy than the original TARP algorithm. When deploying the xTARP method in the Fuzzy Vault construction of Butt et al., the false non-match rate (FNMR) at a security level of 20 bits is reduced from 7.4% to 1.7%.

Keywords: Fingerprint Registration, Reference Point Detection, Biometric Template Protection.

1 Introduction

Reliable reference point detection is an important building block for identification systems and biometric cryptosystems based on fingerprint minutiae. With the help of a reference point, an absolute pre-alignment can be applied for fingerprint registration, thereby compensating variations in the placement (translation) and, if constituted with a direction, rotation of different imprints from the same finger.

The most prominent reference points are the singular points of the orientation field, i.e. core and delta points. Many algorithms for singular point detection have been proposed, e.g. [ZHY01, BG02, NB03]. However, fingerprints of type *arch* do not have any singular points and, therefore, singular point detection alone is not a universal approach. Many publications proposed algorithms for the estimation of generalized singular points, e.g., highest curvature points, that are also present in arch-type fingerprints, e.g. [Ja00, RA00, LJK05, Ig06, GZY16]. An alternative approach is the estimation of a so-called *focal point* [ASJ06, AB08, BA09]. However, most of these universal methods (exceptions being [LJK05] and [BA09]) do not output a direction which could be used to compensate different rotations of the fingerprints.

[1] secunet Security Networks, Mergenthaler Allee 77, Eschborn, Germany, johannes.merkle@secunet.com
[2] secunet Security Networks, Konrad-Zuse-Platz 2, München, Germany, benjamin.tams@secunet.com
[3] secunet Security Networks, Mergenthaler Allee 77, Eschborn, Germany, benjamin.dieckmann@secunet.com
[4] Bundesamt für Sicherheit in der Informationstechnik, Bonn, Germany, ulrike.korte@bsi.bund.de

A promising approach to determine directed reference points was published in [Ta13] and [TMM15]. There, the actual fingerprint's orientation field is aligned with a mathematical orientation field model for fingerprints of type *tented arch*.[5] The alignment between the orientation field of the fingerprint and the orientation field of the mathematical model results from a minimization of a cost function measuring the deviation between these fields in a region around the model's core point. In [Ta13, TMM15, Bu16, Ta16, Ta15], this *Tented Arch Reference Point* (TARP) algorithm was successfully applied in biometric cryptosystems based on the *fuzzy vault scheme*. However, the achieved error rates were still too high for practical applications, which indicates that the TARP algorithm is not yet sufficiently accurate.

We improve the TARP algorithm by several means. Firstly, we integrate an improved orientation field computation. Furthermore, we implement an additional mathematical model for arch-type fingerprints and develop a fusion method to make use of both models. As a result, we obtain an *Extended Tented Arch Reference Point* (xTARP) algorithm which exhibits a considerably higher accuracy than the TARP and other reference point estimation methods. We show the utility of our xTARP algorithm by deploying it for pre-alignment in the Fuzzy Vault construction of [Bu16] (instead of the TARP algorithm).

The structure of this paper is as follows. We give a brief introduction into the TARP method in Section 2 and analyze its potential for improvement in Section 3. In Section 4, we describe our improvements. The resulting xTARP algorithm is evaluated and compared with other methods for singular point detection in Section 5. Finally, we draw conclusions in Section 6.

2 The Tented Arch Reference Point Algorithm

The Tented Arch Reference Point (TARP) algorithm for estimating directed reference points was proposed in [Ta13] and (with slightly improved parameters) in [TMM15] and its implementation in C++ was published under LGPLv3 license.[6]

Fig. 1: For most fingerprints, the ridge flow around the center resembles that of a fingerprint of type tented arch.

[5] Note, that this tented arch model is applied to fingerprints of all types
[6] http://www.stochastik.math.uni-goettingen.de/biometrics/thimble

The basic idea of the TARP algorithm is that, for most fingerprints, the ridge flow in a ring-shaped area around the center resembles that of a fingerprint of type *tented arch*, as illustrated in Figure 1. Therefore, the TARP algorithm attempts to align the orientation field of the fingerprint to a fixed directional field resembling the orientation field of fingerprints of type tented arch. The directional field is derived from a modification of the the *quadratic differential* (QD) model [HHM08], which represents fingerprint orientation fields of arch-type fingerprints by the complex function

$$\psi(z) = \lambda^2(z^2 - R^2), \tag{1}$$

with $\mathrm{Im}(z) > 0$ and real-valued parameters λ, R. The orientation at location (a,b) is defined by $\phi = 0.5\mathrm{Arg}(\psi(a+i\cdot b))$. For fingerprints of type *tented arch*, the model is extended to $\tau(z) = \psi(z)\frac{z^2+d_{\mathrm{core}}^2}{z^2+d_{\mathrm{delta}}^2}$, where $(0,d_{\mathrm{core}})$ and $(0,d_{\mathrm{delta}})$ with $d_{\mathrm{core}} \geq d_{\mathrm{delta}} \geq 0$ are the positions of the core and delta point, respectively, in the model. We refer to [HHM08] for further details and explanations.

By setting $\tau_{\alpha,\beta}(z) = \alpha^{-2}\tau(\alpha z + \beta)$, the TARP algorithm applies translations β and rotations α (with $\alpha, \beta \in \mathbb{C}$ and $|\alpha| = 1$)[7] to the tented arch model $\tau(z)$ to find the best fit with the fingerprint orientation field in a region around the model's core point. In this translated and rotated model, the core point's position is $\gamma_{\alpha,\beta} = \alpha^{-1}(i\cdot d_{\mathrm{core}} - \beta)$. If the fingerprint's orientation field is given by $\{(z_j, v_j)\}$, where the z_j are quantized positions and v_j the orientation at position z_j, its fit with the model $\tau_{\alpha,\beta}(z)$ is evaluated using the cost function

$$\kappa(\alpha,\beta) = \sum_j w_{\alpha,\beta}(z_j)\left|\frac{\tau_{\alpha,\beta}(z_j)}{|\tau_{\alpha,\beta}(z_j)|} - v_j\right|; \tag{2}$$

here, $w_{\alpha,\beta}(z) = \exp\left(-\frac{(|z-\gamma_{\alpha,\beta}-\rho|)^2}{2\sigma^2}\right)$ is a weight function resulting from a Gaussian function with deviation σ rotated at distance ρ around the core $\gamma_{\alpha,\beta}$. [8] The reference point is defined as the core point $\gamma_{\alpha,\beta}$ for those α,β that minimize the cost function; its direction is given by the model's rotation $\mathrm{Arg}(\alpha)$.

In order to minimize the cost function (2) over α and β, a two-step approach is performed. In an initial search step, for all positions in a rectangular grid of width g and for $\alpha = 0$, the corresponding translation β is computed and the cost function $\kappa(\alpha,\beta)$ is evaluated. Starting with the value (α,β) from the initial search resulting in the smallest $\kappa(\alpha,\beta)$, a steepest descent algorithm is performed in which the cost function is further minimized. In the steepest descent, the rotation α and the translation β are refined alternatively, until the value of $\kappa(\alpha,\beta)$ converges to a (local) minimum. If the steepest descent does not result in a reference point inside the fingerprint's foreground, it is repeated starting with the next best candidate from the initial search. The iteration continues until a reference point on the fingerprint's foreground is found or the steepest descent has failed $n_{\max}$ times.

[7] Recall that when representing numbers in the complex plane, addition translates to vector addition and multiplication results in addition of the arguments (angles) and multiplication of the absolute values (lengths).

[8] The distance ρ of the Gaussian function's center from the core point results in a ring-like shape (with smooth borders) of the weight function so that the region directly at the core point, where the orientation field computation is often unreliable, is less weighted.

A visualization of the TARP model and its parameters can be found in [TMM15]. In [TMM15], the parameters $\lambda = 1.81, d_{\text{core}} = 160, d_{\text{delta}} = 22, R = 175, \rho = 45, \sigma = 12, g = 7$ and $n_{\max} = 20$ were chosen based on a training on the FVC2000 DB2b database comprising 80 fingerprints.

3 Analysis of the Improvement Potential

We identified the following potential sources of inaccuracies of the TARP algorithm.

- **Segmentation.** The TARP algorithm performs segmentation by a combination of Otsu thresholding [LO79] and Graham scan [Gr72]. This relatively simple approach sometimes fails, e.g., when the imprints of the first and the second finger pad are not clearly separated.
- **Orientation field.** The TARP algorithm computes orientation fields by local intensity gradients without any post-processing. Therefore, errors in the orientation fields may contribute considerably to the inaccuracies of the reference point detection.
- **Model.** As visible in Figure 1, the tented arch model fits well for most types of fingerprints but not that good for arch-type fingerprints.
- **Minimization.** The steepest descent (gradient) method may yield merely a local (but not global) minimum of the cost function.

In order to focus our effort to improve the TARP algorithm on those aspects that promise the highest gain in accuracy, we analyzed the impact of each of these factors to the accuracy of the reference points. For that purpose, we manually marked ground truth data for the segmentation and orientation fields of 273 optical sensor fingerprints of right index fingers from the MCYT database [Or03] for which the TARP algorithm showed relatively poor performance. Furthermore, we modified the TARP algorithm so that it can be fed with externally generated segmentation maps and orientation fields. Then, we computed the reference points for these fingerprints with and without the ground truth segmentation maps and/or orientation fields and evaluated the accuracy of the alignments resulting from these reference points, using the metric described in Section 5. In order to investigate the third factor, we manually determined the type of all imprints of right index fingers of the MCYT database and, for each type, evaluated the accuracy of the alignments resulting from the reference points computed with the TARP algorithm. Furthermore, we modified the TARP implementation so that the minimum of the cost function is determined by exhaustive search, and evaluated the accuracy in comparison to the original method (steepest descent).

The results in Figure 2 clearly show that the errors in the computation of orientation fields has much more impact on the accuracy than incorrect segmentation, and that the TARP algorithm performs much worse on arch-type fingerprints than on other types of fingerprints. Compared to these factors, the minimization of the cost function showed slightly less potential for improvement. Therefore, we decided to improve the TARP algorithm by deploying a better orientation field computation method and by extending the tented arch model to arch type fingerprints.

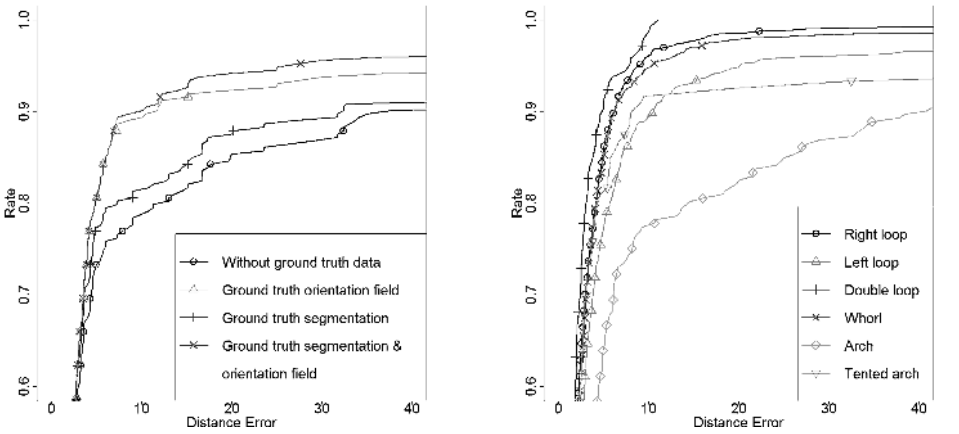


Fig. 2: Evaluation of the dependence of the accuracy of the TARP algorithm on errors in segmentation and/or orientation field computation (left hand side) and on the type of fingerprint. The diagrams show the ECDF of distance errors on two different sets of fingerprints from the MCYT database.

4 Improvements

4.1 Improved Orientation Field Computation

In order to improve the orientation field computation, we deployed the algorithm described in [LP08], which applies a Markov Random Field to improve the initial orientation field obtained by a gradient approach. The Markov Random Field is constructed from two components: a component based on a global mixture model obtained from training on real fingerprints and a component enforcing pairwise consistency of neighboring image blocks. We used an implementation of this algorithm contained in the FingerJetFX open source minutiae extractor[9].

In [TMM15], the parameters of the TARP method were chosen based on a training on the FVC2000 DB2b database comprising 80 fingerprints. Since this training set was quite small and the improved orientation field computation may favor different parameters, we conducted a new training on a large training set comprising 2736 fingerprints of 228 right index fingers from the MCYT database [Or03]. Precisely, we deployed the TARP algorithm in the Fuzzy Vault implementation of [Bu16] and evaluated the FNMR for a degree $k = 6$ of the secret polynomials (which gives approximately 20 bits of security). We found the highest accuracy for $\lambda = 1.3, d_{\text{core}} = 340, d_{\text{delta}} = 32, R = 185, \rho = 28, \sigma = 6$. The best trade-off between accuracy and processing time was found for $g = 7$ and $n_{\max} = 5$. Note that the parameters $d_{\text{core}}, d_{\text{delta}}, R, \rho, \sigma, g$ apply for 500 dpi images and need to be linearly scaled for other resolutions.

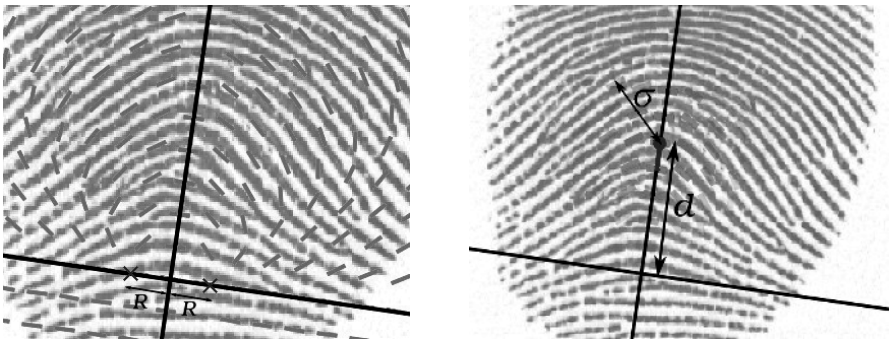
[9] https://github.com/FingerJetFXOSE/FingerJetFXOSE

4.2 Developing a Modified Algorithm for Arches (ARP)

Furthermore, we designed and implemented an Arch Reference Point (ARP) detection algorithm based on the quadratic differential model $\psi(z)$ for arch-type fingerprints specified in (1). Analogously to the TARP algorithm, it tries to find a translation $\beta \in \mathbb{C}$ and a rotation $\alpha \in \mathbb{C}$ (with $|\alpha| = 1$) so that the correspondingly transformed model $\psi_{\alpha,\beta}(z) = \alpha^{-2}\psi(\alpha z + \beta)$, fits best with the fingerprint's actual orientation field in the area around a reference point. In the original model $\psi(z)$, this reference point is set to $(0,d)$, and after transformation by α,β its position is $\overline{\gamma}_{\alpha,\beta} = \alpha^{-1}(i \cdot d - \beta)$. The fit of the fingerprint's orientation field $\{(z_j, v_j)\}$ with the model $\psi_{\alpha,\beta}(z)$ is evaluated using a cost function

$$\overline{\kappa}(\alpha,\beta) = \sum_j \overline{w}_{\alpha,\beta}(z_j) \left| \frac{\psi_{\alpha,\beta}(z_j)}{|\psi_{\alpha,\beta}(z_j)|} - v_j \right|, \tag{3}$$

where $\overline{w}_{\alpha,\beta}(z) = \exp\left(-\frac{(|z-\overline{\gamma}_{\alpha,\beta}|)^2}{2\sigma^2}\right)$ is a weight function resulting from a two-dimensional Gaussian function with deviation σ and center $\overline{\gamma}_{\alpha,\beta}$. [10] A visualization of both the ARP model and the weight function is shown in Figure 3.



(a) (b)

Fig. 3: Visualization of the ARP model (a) and the weight function (b) and the corresponding parameters. The parameter λ is not visualized; it controls how stretched the ARP model is.

For the minimization of the cost function (3) over α and β, the same two-step method as in the TARP algorithm, i.e. an initial search over a rectangular grid of width g followed by a steepest descent (see Section 2), is applied, using the parameters $m, r, \alpha_{\max}$ (width of the grid, number of rotations and maximum rotation angle used in the initial search) and $n_{\max}$ (maximum number of attempts for the steepest descent). After minimizing $\overline{\kappa}(\alpha,\beta)$ over α,β, the algorithm outputs the reference point $\overline{\gamma}_{\alpha,\beta}$ with orientation $\mathrm{Arg}(\alpha)$.

Analogously to the training of the improved TARP method, we optimized the parameters of the ARP algorithm by deploying it in the Fuzzy Vault implementation of [Bu16] and

[10] The weight function $\overline{w}_{\alpha,\beta}$ has its maximum at the reference point, in contrast to the weight function $w_{\alpha,\beta}$ used in the TARP model, which has its maximum on a circle with radius ρ around the core point. (see Footnote 8)

evaluating the false reject rate (FRR) at $k = 6$ for a training set of 393 arch-type fingerprints (captured with an optical sensor) of 33 left and right index fingers from the MCYT database [Or03]. The best recognition accuracy was achieved for $\lambda = 1.5, d = 80, R = 20, \sigma = 27, g = 5, n_{\max} = 5$. Analogously to the parameters of the TARP method, the values of the parameters d, R, σ, g apply for 500 dpi images and need to be linearly scaled for other resolutions.

4.3 Implementing xTARP by Fusion of TARP and ARP

In order to combine TARP and ARP in an Extended Tented Arch Reference Point (xTARP) method, we needed a rule to decide which of the two points to choose. We could have tried to use a classifier to detect arch-type fingerprints, but as it turned out in our evaluation, the ARP method works well not only on this fingerprint class. Since the minimized values κ and $\overline{\kappa}$ of the cost functions (2) and (3) of the TARP and the ARP algorithm, respectively, indicate how good the fingerprint's orientation field fits with the respective QD model, we decided to implement a classifier using the values κ and $\overline{\kappa}$ as input. For a training set of 4920 fingerprints of 410 left and right index fingers from the MCYT database [Or03], we computed reference points with both algorithms and used these for pre-alignment in the Fuzzy Vault implementation of [Bu16]. Then, we selected as training classes two subsets of fingerprints, for which one of the two methods (TARP or ARP, respectively) yields less than 8 genuine points in the unlocking set but the other method results in at least 8 genuine points.

Since a genuine comparison will already yield a poor result if one of the two reference points is inaccurate, the larger one of the cost function values of the reference and the query fingerprint should be more indicative for the genuine score than the lower one. Therefore, as input data for the training we computed, for each genuine pair, the maximum $\kappa_{\max} = \max(\kappa_{\text{ref}}, \kappa_{\text{que}})$ of the TARP cost values of reference and query fingerprint and, analogously, the maximum $\overline{\kappa}_{\max} = \max(\overline{\kappa}_{\text{ref}}, \overline{\kappa}_{\text{que}})$ of both cost values of the ARP algorithm. For the two classes and the input vector $(\kappa_{\max}^{[i]}, \overline{\kappa}_{\max}^{[i]})$ we optimized a linear discriminant function, starting with logistic regression and optimizing the coefficients by evaluating the resulting FRR of the Fuzzy Vault with $k = 6$ for the complete training set. We obtained the best results for the linear discriminant function $y = 0.6 \cdot \overline{\kappa} - \kappa + 10$, i.e., when selecting the ARP point if and only if $\kappa > 0.6 \cdot \overline{\kappa} + 10$. This classification rule was implemented into our xTARP algorithm. We also investigated classifiers based on support vector machines with various kernels but did not obtain significantly better results.

5 Experiments

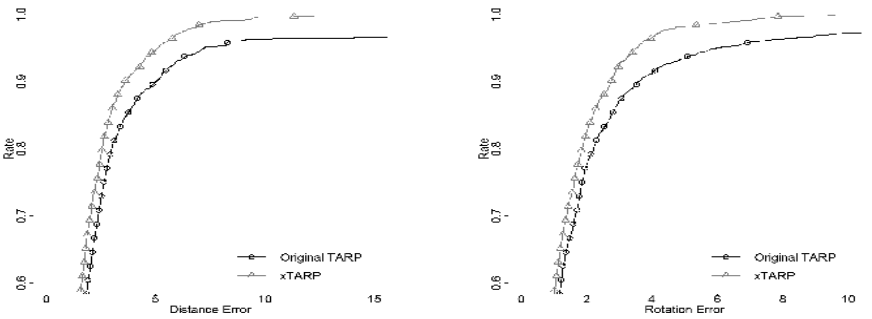
We performed experiments to evaluate the accuracy of the xTARP algorithm and to compare it with the original TARP algorithm from [TMM15] and other methods for reference point detection.

Since the ARP and TARP reference points cannot be visually identified by humans, like core or delta points, we cannot measure the accuracy of the reference points with respect to ground-truth data. In order to overcome this problem and to allow a fair comparison with other publications, we compute, for each finger, an approximation of the "true reference point" by means of the median of the computed reference points. Precisely, for each fingerprint, we manually determined the affine transformations (rotation and translation) by which it is aligned with the other fingerprints of the same finger.[11] Using these transformations, we projected the reference points of the other fingerprints into this fingerprint. From all projected reference points (with accordingly rotated orientations) and the reference point of this finger, we computed the median of the positions and orientations as approximated "true reference point".

For a reference point (x, y, θ), we measured the distance error *DE* as the Euclidean distance to the approximated "true reference point" $(\bar{x}, \bar{y}, \bar{\theta})$ of the fingerprint, i.e. as $DE = \sqrt{(x-\bar{x})^2 + (y-\bar{y})^2}$, and the rotation error *RE* as the absolute value of the smaller angle (in degrees) between their directions, i.e. as $RE = \min(\delta_\theta, 360 - \delta_\theta)$, where δ_θ is the representative of $(\theta - \bar{\theta})$ mod 360 in the interval $[0, 360)$.

5.1 Evaluation of Optimizations

First, we evaluated the effectiveness of our improvements by comparing the accuracy of our xTARP algorithm and the original TARP method from [TMM15]. For this evaluation, we used as a test set the optical sensor fingerprints of right index fingers of the first 100 subjects in the MCYT database [Or03]. This test set is disjoint to the training sets used for optimization of parameters and fusion (Section 4), and it comprises 1200 fingerprints (12 per finger) of relatively high quality taken with the sensor *UareU* from Digital Persona at 500 dpi and stored in uncompressed image files of 256×400 pixels. The results show that both distance errors and rotation errors are considerably reduced (Figure 4).



(a) (b)

Fig. 4: ECDF of distance errors (a) and rotation errors (b) of the original TARP method (black) and our xTARP algorithm (red).

[11] In cases where non-linear distortions did not allow an accurate global alignment, we chose a transformation that aligns the central region of the finger pad where the reference points are expected to be located.

Whereas the original TARP algorithm fails (i.e., does not find a reference point) for 0.5% of the images, the xTARP algorithm processes all of them successfully. The optimized TARP algorithm (as described in Section 2) already achieves error rates close to that of the xTARP method, but it fails for 0.67% of the images.

5.2 Comparison with Other Approaches

In order to compare our xTARP algorithm with other methods for reference point detection, we evaluated it on FVC2000 DB2a as well. This data base comprises 800 fingerprints from right index fingers of 100 subjects (8 per finger), acquired by a low cost capacitive sensor at 500 dpi and stored in image files with loss-less compression and 256×364 pixels.

Table 1 compares the accuracy of the xTARP method[12] with other state-of-the-art methods for reference point detection. Note, that the methods of [LJK05, Ig06, AB08, GZY16] do not determine any orientation of the reference point (indicated by "n.a.").

Method	$DE < 5$	$DE < 10$	$DE < 20$	$RE < 5$	$RE < 11.25$	$RE < 22.5$	Fail
xTARP	612	734	763	610	752	790	5
[GZY16]	569	719	784	n.a.	n.a.	n.a.	0
[BA09]	n.a.	668	769	n.a.	521	657	0
[AB08]	285	640	763	n.a.	n.a.	n.a.	1
[Ig06]	n.a.	712	753	n.a.	n.a.	n.a.	0
[LZH06]	n.a.	654	745	n.a.	690	737	9
[LJK05]	n.a.	659	749	n.a.	n.a.	n.a.	13

Tab. 1: Cumulative statistics of the distance errors (DE), rotation errors (RE), and number of failures (no reference point is output) of our xTARP method and other reference point detection methods. An entry "n.a." means that the corresponding value is not provided in the referenced publications or (in the case of RE) that the method does not compute any orientation of the reference point.

As shown in Table 1, the xTARP method outperforms all other methods with respect to distance errors for up to 10 pixels. Furthermore, xTARP is by far the most accurate method with respect to rotation errors.

On the other hand, our improvements of the TARP algorithm are much less effective for the FVC2000 DB2a database as they are for the MCYT database, i.e., the original TARP method already exhibits a similar accuracy as the xTARP method. The reasons of this finding are yet to be analyzed.

5.3 Application in a Fuzzy Vault for Fingerprints

In order to prove the utility of the xTARP method, we applied it for the pre-alignment of the fingerprints in the Fuzzy Vault construction of [Bu16]. Precisely, we deployed the

[12] Based on a training on the FVC 2000 DB2a, we slightly adapted the linear discriminant function for the fusion (see Section 4.3).

variant which uses minutiae's positions and orientations but not their type and applied the same parameters as in [Bu16] for the degree k of the (monic) secret polynomial from 5 to 8. Our evaluation shows that the false match rate (FMR) does not depend on the pre-alignment method, which is quite plausible because the concept of an exact alignment does not make sense for impostor verifications. Therefore, the estimates from [Bu16] for the security level and FMR also apply for an alignment with our improved methods.

Due to the fusion of two different models (ARP and TARP) by the xTARP method, it can happen that a different model is used for the reference fingerprint as for the query fingerprint. Such an inconsistent application of the models almost always results in a non-match, because the location of TARP and ARP points are typically different. To overcome this problem, we store a status bit with the reference template, indicating which of the two methods were used during enrolment, and use the same method for verification. By this approach, we operate the xTARP method in a *stateful mode*, in contrast to the *stateless mode*, where the decision between ARP and TARP is taken for each fingerprint independently. Another advantage of the stateful mode is that only one reference point is computed for the query fingerprint and, thus, the processing time of the verification is reduced. For the stateful mode, we found a slightly different linear discriminant function $y = 0.8 \cdot \overline{\kappa} - \kappa - 28$ to be optimal for fusion (see Section 4.3).

For the various stages of development of the xTARP method, we evaluated the FNMR[13] on the same test set as in [Bu16] (right index fingers of the first 100 subjects of the MCYT data base). The results in Table 2 show that the FNMR is already greatly reduced for the TARP method with the improved orientation field estimation, and even further by the xTARP method. For the xTARP method, the storage of a status bit along the template indicating whether a TARP or an ARP reference point has been used for enrolment (stateful mode) further improves the error rates.

	Method	$k=5$	$k=6$	$k=7$	$k=8$
FNMR	Original TARP	6.0%	7.4%	9.6%	13.1%
	Improved TARP	1.6%	2.8%	4.8%	7.6%
	ARP	2.0%	3.5%	5.9%	9.6%
	xTARP (stateless)	0.5%	1.7%	3.8%	6.6%
	xTARP (stateful)	0.5%	1.5%	3.4%	6.3%
FMR	all	1.9%	0.3%	0.04%	0%
Security (bits)	all	16.5	20	24	27

Tab. 2: Error rates of the Fuzzy Vault construction of [Bu16] when using different variants of the xTARP and TARP method for pre-alignment.

The FNMR can be even further reduced to 0.2%, 0.8%, 2.2% and 4.6%, respectively, if during verification, in case of a non-match, a second attempt is conducted with a reference point computed with the orientation field estimation from the original TARP algorithm.

[13] In this Fuzzy Vault construction, failures of the reference point detection also contribute to the FNMR.

6 Conclusions

We have greatly improved the accuracy of the TARP reference point detection method by deploying a better algorithm for orientation field estimation and by complementing the tented arch model with a model for plain arches. When using the resulting xTARP method in a biometric cryptosystem, the FNMR for a security level of 20 bits drops from 7% to 1.7%.

Nevertheless, we still see significant potential for further improvements: the xTARP algorithm tends to be inaccurate for fingerprints where the true reference point is close to the edge of the foreground, which is often the case for the FVC2000 DB2a data base. In these regions, the cost function is partially evaluated on the background, where the orientations are constant or randomly chosen (we tried both options without obtaining significantly different results), which increases the cost value and may imply that the minimum is found in a different region. Therefore, we suspect that a better performance can be achieved by limiting the cost function to the foreground. However, for the deepest descent method to work, the cost function has to be smooth.

Furthermore, due to the combination of two different fingerprint models, the computational effort is quite large and should be improved. Currently, the xTARP algorithm takes more than 1.5 seconds in stateless mode and 0.7 seconds in stateful mode. Since the algorithm has not been optimized yet with respect to computational complexity, we believe that there is great potential to tab. For instance, the computation of the cost function is based on convolution and, thus, could be considerably sped up using Fast Fourier Transformation.

References

[AB08] Areekul, Vutipong; Boonchaiseree, Natthawat: Fast focal point localization algorithm for fingerprint registration. In: Industrial Electronics and Applications, 2008. ICIEA 2008. 3rd IEEE Conference on. IEEE, pp. 2089–2094, 2008.

[ASJ06] Areekul, Vutipong; Suppasriwasuseth, Kittiwat; Jirachawang, Suksan: The New Focal Point Localization Algorithm for Fingerprint Registration. In: 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China. pp. 497–500, 2006.

[BA09] Boonchaiseree, Natthawat; Areekul, Vutipong: Focal Point Detection Based on Half Concentric Lens Model for Singular Point Extraction in Fingerprint. In: Advances in Biometrics, Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings. pp. 637–646, 2009.

[BG02] Bazen, Asker M.; Gerez, Sabih H.: Systematic Methods for the Computation of the Directional Fields and Singular Points of Fingerprints. IEEE Trans. Pattern Anal. Mach. Intell., 24(7):905–919, 2002.

[Bu16] Butt, Moazzam et al.: Correlation-resistant fuzzy vault for fingerprints. In: Proc. of Sicherheit 2016. volume 256 of LNI. GI, pp. 125–136, 2016.

[Gr72] Graham, Ronald L.: An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. Inf. Process. Lett., 1(4):132–133, 1972.

[GZY16] Guo, Xifeng; Zhu, En; Yin, Jianping: A fast and accurate method for detecting fingerprint reference point. Neural Computing and Applications, pp. 1–11, 2016.

[HHM08] Huckemann, Stephan; Hotz, Thomas; Munk, Axel: Global Models for the Orientation Field of Fingerprints: An Approach Based on Quadratic Differentials. IEEE Trans. Pattern Anal. Mach. Intell., 30(9):1507–1519, 2008.

[Ig06] Ignatenko, Tanya et al.: Reference point detection for improved fingerprint matching. In: Security, Steganography, and Watermarking of Multimedia Contents VIII, San Jose, CA, USA, January 15, 2006. p. 60720G, 2006.

[Ja00] Jain, Anil K. et al.: Filterbank-based fingerprint matching. IEEE Trans. Image Processing, 9(5):846–859, 2000.

[LJK05] Liu, Manhua; Jiang, Xudong; Kot, Alex ChiChung: Fingerprint Reference-Point Detection. EURASIP J. Adv. Sig. Proc., 2005(4):498–509, 2005.

[LO79] Level Otsu, N: A threshold selection method from gray-level histogram. IEEE Transactions on Systems, Man and Cybernetics, 9(1):62–66, 1979.

[LP08] Lee, Kuang-chih; Prabhakar, Salil: Probabilistic orientation field estimation for fingerprint enhancement and verification. In: Biometrics Symposium, 2008. BSYM'08. IEEE, pp. 41–46, 2008.

[LZH06] Liu, Tong; Zhang, Chao; Hao, Pengwei: Fingerprint Reference Point Detection Based on Local Axial Symmetry. In: 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China. pp. 1050–1053, 2006.

[NB03] Nilsson, Kenneth; Bigün, Josef: Localization of corresponding points in fingerprints by complex filtering. Pattern Recognition Letters, 24(13):2135–2144, 2003.

[Or03] Ortega-Garcia, J. et al.: MCYT baseline corpus: a bimodal biometric database. IEE Proc. on Vision, Image and Signal Processing, 150(6):395–401, 2003.

[RA00] Rerkrai, Krisakorn; Areekul, Vutipong: A New Reference Point for Fingerprint Recognition. In: Proceedings of the 2000 International Conference on Image Processing, ICIP 2000, Vancouver, BC, Canada, September 10-13, 2000. pp. 499–502, 2000.

[Ta13] Tams, Benjamin: Absolute Fingerprint Pre-Alignment in Minutiae-Based Cryptosystems. In: Proc. of BIOSIG 2013. volume 212 of LNI. GI, pp. 75–86, 2013.

[Ta15] Tams, Benjamin et al.: Improved Fuzzy Vault Scheme for Alignment-Free Fingerprint Features. In: Proc. of BIOSIG 2015. volume 245 of LNI. GI, 2015.

[Ta16] Tams, Benjamin: Unlinkable minutiae-based fuzzy vault for multiple fingerprints. IET Biometrics, 5(3):170–180, 2016.

[TMM15] Tams, B.; Mihăilescu, P.; Munk, A.: Security Considerations in Minutiae-based Fuzzy Vaults. IEEE Trans. Inf. Forensics Security, 10(5):985–998, 2015.

[ZHY01] Zhang, Qinzhi; Huang, Kai; Yan, Hong: Fingerprint Classification Based on Extraction and Analysis of Singularities and Pseudoridges. In: Visualisation 2001, Selected Papers from the Pan-Sydney Area Workshop on Visual Information Processing, VIP2001. pp. 83–87, 2001.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 83

Fingerprint Template Ageing vs. Template Changes Revisited

Simon Kirchgasser[1], Andreas Uhl[2]

Abstract: This study investigates the impact of "ghost" fingerprint and minutiae information in 4 year time-span separated fingerprint datasets. A high amount of ghost fingerprints within the data, eventually a source for differences in acquisition conditions, might be responsible for recently reported template ageing effects. According to that, various experiments have been performed to get rid of this problematic image content and to compare the corresponding matching results to the performance figures using the non altered imprints. The analysis with respect to detected increased error rates exhibits very similar effects for all considered methods no matter if ghost fingerprint information is removed or not. Thus, ghost fingerprints are not responsible for the observed effects.

Keywords: Fingerprint Recognition, Template Ageing, Quality, Ghost Fingerprints.

1 Introduction

The ISO/IEC biometric testing standard ISO/IEC 19795-1 reports that "Longer time intervals generally make it more difficult to match samples to templates due to the phenomenon known as template ageing" [Ma05]. The standard then defines "template-ageing" as an "increase in error rates caused by time-related changes in the biometric pattern, its presentation, and the sensor". Apart from time-related changes various other reasons can cause performance degradations in fingerprint (FP) recognition as well. The most prominent ones are the usage of different sensors and sensor types, alternation in ambient conditions (e.g. changes in the illumination set-up), differences in the acquisition protocol like variability in sensor plates' cleaning, weather conditions, or various skin diseases as reported in [Dr12].

Considering the high number of potential reasons for FP recognition accuracy degradations, we investigate a different (i.e. not time-related) explanation for the recently postulated template ageing effects on time separated data [KU17a, KU17b] in this work. In [FCM17] it is confirmed that a) FP images can be designed which include the biometric minutiae information of at least 2 fingers and b) that such imprints cause serious troubles during the recognition process using state-of-the-art implementations. In Figure 2, displaying example imprints of the datasets used in [KU17a, KU17b], it is easy to find minutiae information in the background, which clearly do not belong to the acquired finger in the region of interest (ROI). This additional information, a so called "ghost" FP, can be found very frequently in the considered datasets. It is rather obvious that a ghost FP would not cause any decrease in the quality measure analysis as performed in [KU17a, KU17b]. Further, the presence of ghost FP was discussed as a complicating factor during FP segmentation in [THG16, WTG07, Zh06] and most importantly, the detailed observation of our considered imprints revealed that background information (i.e. ghost FP) is not always present in each image of the used data. There are images which contain identical ghost

[1] University of Salzburg, Jakob-Haringer-Str. 2, 5020 Salzburg, AUSTRIA, skirch@cosy.sbg.ac.at
[2] University of Salzburg, Jakob-Haringer-Str. 2, 5020 Salzburg, AUSTRIA, uhl@cosy.sbg.ac.at

FPs each time an imprint of the same finger is acquired. But, there are also FP images available where no such information can be retrieved. This alteration in the presence of ghost FPs actually leads to template changes, which could cause changing error rates and thus could be made responsible for template ageing effects. However, it is not correct that these changes can be classified as being time-related. A varying presence of ghost FPs is caused by acquisition protocol variations, i.e. the definition when the sensor surface is being cleaned. Of course, acquisition protocols differing with respect to this property could be used in two sessions without any time separation in-between. Thus, if our experiments reveal that ghost FPs cause the observed effects, template ageing is not the reason but a time-unrelated template change effect.

The rest of this paper is organised as follows: In Section 2, we review the current state of the art on the relation of fingerprint recognition and ageing. The experimental setup, i.e. the used FP recognition SDKs, datasets and a detailed discussion on the used experimental methodology will be presented in Section 3. The subsequently performed experiments and corresponding results are analysed in Section 4, before concluding this study in Section 5.

2 Fingerprint Recognition and Ageing

The biological reason for FP ageing is the loss of collagen [Mo07]. This structural protein ensures that the human skins' fibrous tissue is resilient during time. Even though, it is possible to measure skin ageing. The most prominent methods are the usage of high-frequency skin ultrasonography, prophilometry and skin micro-relief descriptors [BG04]. Furthermore it is even possible to describe skin topography changes from capacity images by analysing the 3D profile. This analysis reveals the introduction of wrinkles and a cell enlargement caused by the biological ageing process [GJ98]. Uchida et al. [Uc96] quantify skin ageing by analysing the 3D profile of subjects aged 20-60 using 2D DFT features (assessing skin ridges) resulting in less high frequency components for elder people - but also wide scattering. But there are also more recent studies which focus on the ageing behaviour of latent FP, being of high importance in crime scene analysis, looking into biological aspects in more detail. First the FP information of the various test subjects was deposited at e.g. glass or synthetic material. The particular biometric traits were acquired after some period exhibiting different time-spans. In [PPP10] the relationship of these latent FPs, their corresponding time-spans and biological degradations during the specified time period was investigated. Apart from classical examination methods like morphological and structural approaches, biochemical and DNA based tests have been used as well to measure FP degradations. The investigations revealed that for example the blood groups do have an influence on the degradation. It seems that people exhibiting blood group B are slightly more resistant to biochemical ageing influences. Of course those results are more important for forensic datasets, but small biochemical variations could also lead to degradations which can influence the recognition process. The authors of this particular study used 800 FP images for the performed experiments. Further specifications on the used analysis tools, e.g. microscope and DNA extraction process, may be looked up in [PPP10].
Another biological aspect was investigated in [Me13], using chromatic white light sensors to study latent long-term FP ageing. The authors state that an image contrast loss can

be observed over time, considering imprints of 40 volunteers. The corresponding images have been acquired at three different locations independently and were compared during the experiments based on four different research goals. The results revealed a high number of variance among the different time series of user's FP images. The authors concluded that the reason for this observation might be a different biochemical composition of the imprints.

2.1 Fingerprint Age Group Analysis

Focusing on the aspect of human ageing it is natural that studies have been performed, which investigate the influence of different subject age groups in FP datasets on recognition performance. In [SE05] it was shown that older age groups exhibit a worse performance in terms of FP quality and recognition performance. This conclusion was achieved by analysing the relationship between FP's moisture content and the volunteer's age using a one-way analysis of variance (ANOVA) and the Pearson correlation coefficient. The corresponding database contains images of 79 people (age group from 18-25) and imprints of 60 people (age group 62+). In total 948 images of age group 18-25 and 720 of the second age group are included in this dataset. Of each volunteer, 3 images of each index finger (left and right hand) have been acquired. This database was reused in subsequent research [ME06], where the authors focused on minutiae point based analysis. This resulted in the conclusion that elderly people exhibit a higher number of minutiae points, but the biometric quality (using NIST Fingerprint Image Quality algorithm[3]) displayed a degradation compared to the younger age group. Finally in [Mo07] this investigation was extended once more. The dataset was expanded by two additional age groups (26-39 and 40-64). The authors could confirm the results stated by [SE05] that older age groups are displaying a worse performance in terms of FP quality and recognition.

In [UW09] a similar study was performed, but the core aspect of this research was the consideration of a different dataset exhibiting very young people as well. Not only age groups of volunteers older than 19 years have been taken into account, but also the age group from 3 to 18 years. According to this aspect two different sub-datasets have been acquired: One containing the adult biometric templates (172 in total) and one displaying the young volunteers' images (498 in total). Further specific information on the volunteers can be looked up in [UW09]. Additionally it must be mentioned that the acquisition was done by the use of a optical scanning device (a HP 3500c flatbed scanner) with 500 dpi resolution, capturing the full hand. Data analysis was done by the usage of 5 different (hand-)geometric and texture-based methodologies, including FP minutiae, eigenfingers, geometric and shape based approaches. The interested reader is referred to [RF05] (eigenfingers) and [JRP99] (geometric methods) for more detailed information on those techniques.

The final results concerning the recognition performance are based on three different age groups. These groups have been selected as subsets of the previously introduced adult and children datasets: The first group is called young group and contains all images of children who are between 3 and 10 years old, the second one (youth group) includes the imprints of all volunteers whose age is between 11 and 18 years, and finally the adult group (19+ years). In most performed experiments it can be observed that kids' FP performance suffers compared to adults recognition performance [UW09].

[3] https://www.nist.gov/programs-projects/biometric-quality-homepage

To cope with different age groups and effects which are introduced by the usage of data exhibiting such variability some studies have been performed as well. In [Go11] an isotropic rescaling method was used on children data to improve the recognition performance from $11-14\%$ to $5-6\%$ equal error rate (EER). The experiments were done on imprints, whose feature extraction and matching procedure was improved by analysing the FP's shape and the application of some rescaling approach.

2.2 Fingerprint Ageing Analysis (FP Template Ageing)

Ageing effects in human FP recognition been a topic in research since Galton's first study on the permanence of FPs [Ga92]. In all papers discussed subsequently, increased error rates have been reported for time-separated data. Time intervals of 10 to 30 years have been studied in [ABI05] using a dataset provided by the German federal criminal police office (BKA, i.e. forensic FPs). The authors reported a lower recognition accuracy when the time interval is increased. Further, [RJK07] performed experiments on the so called Korea Fingerprint Recognition Interoperability Alliance (KFRIA) database acquired with three different commercial sensors (2 optical and 1 capacitive sensor type). This dataset exhibits a time span of 1 year between acquisition sessions, which is quite a short time gap, but despite this fact the authors have been able to report an EER increase using three different sensors. The EER of the second acquisition's data was about two times higher than the EER of the corresponding imprint of the first acquisition.
Similar to these results of [ABI05, RJK07], a degradation of different FP matching performance figures (e.g. equal error rate (EER)) was observed on data acquired by a flatbed scanner [UW13], where in particular a decrease of genuine scores was detected for a time separation of 5 years (the genuine scores revealed a decrease of roughly 33% and a 2-4 times lowered EER performance is found). These observations were confirmed on a further massive forensic FP dataset including time-spans up to 7 years in [YJ15] as well. Similar to the detected genuine score degradation on 2D FP data discussed so far, it was possible to observe a decrease in matching and recognition performance using some 3D finger range data which were acquired by covering only a time span of 16 weeks between sessions [WF05]. In [KU16], the presence of similar effects are confirmed on time-separated FP data acquired by off-the-shelf commercial FP scanners by analysing user-group specific effects which are known as the "Doddington Zoo" concept [Do98]. Further investigations on the same data [KU17a, KU17b] revealed very similar effects with respect to decreased recognition performance on time separated data as reported by [UW13, YJ15].
However, most studies done on time-separated FP data have not performed experiments to reveal the reasons for decreased recognition accuracy in detail. In fact, it does not suffice to describe increased error rates on time separated data to have observed a template ageing effect. To be compliant with the definition, time-related changes have to introduce the observed effects, while the sole employment of time-separated data does not automatically imply template ageing being present in case of higher errors (as these might be caused by non-time-related changes). Only few of the studies on time-separated FP data [KU17a, KU17b, YJ15] try to explain why the observed effects occur. The very extensive covariate-fit analysis model in [YJ15] revealed that differences in image quality explain the observed increased errors better as time-related changes. In [KU17a, KU17b] the analysis did not indicate that FP biometric quality decrease can be made responsible for the

claimed template ageing effects. However, these studies unfortunately did not employ the identical experimental and statistical set-up and thus do not even fully clarify the contribution of FP quality to the observed effects, as the results contradict each other.
A potential generic approach to cope with FP template ageing effects is the usage of template update techniques, which have been investigated for example by [KB09]. The authors of this particular study used an adaptive feature set introduced by an algorithm allowing to reduce intra-personal variabilities over time. Similar to this approach there is more recent work focusing on self-updating algorithms [Ma12]. The mentioned update methods provide a path-based clustering setup to enhance the initial template selection before starting the update process on the one hand. On the other hand an improved adaption of the recognition system's threshold is ensured as well in case high environmental variability is measured.

3 Experimental Setup

The experiments have been conducted using two minutiae based FP recognition SDKs: the *NIST Biometric Image Software (NBIS)* and the Neurotechnology *VeriFinger SDK (NEURO)*. The first one (release 5.0.0) has been implemented by the National Institute of Standards and Technology (NIST)[4]. The second recognition approach (release 9.0) was developed by the Lithuanian company Neurotechnology[5].
According to the study purpose we are using datasets already analysed earlier [Ki16, KU17a, KU17b]. The data has been acquired at the Center for Biometrics and Security Research (CBSR) at the Chinese Academy of Sciences, Institute of Automation (CASIA) in 2009 and 2013. The imprints from 2009 are a subset of the publicly available CASIA fingerprint database V5[6]. Using an U.are.U 4000 scanner (produced by DigitalPersona), images of both forefingers and second fingers of 49 volunteers are stored in dataset "CASIA 2009", which will be denote by *A*. In total 980 fingerprint images are available, 5 imprints of each finger. The same acquisition process was repeated four years later to create the "CASIA 2013" database, which includes 5 independent subsets in total. Each subset contains again 980 images of the same volunteers. The main difference among the subsets is the usage of various sensors, among them optical and capacitive fingerprint sensors. They are denoted as *B*1-*B*5. Apart from the "single" datasets containing only imprints of 2009 or 2013 independently, it was necessary to combine the imprints of both years to get so called "crossed", i.e. time-separated, datasets *C*1-*C*5. In each of these crossed sets the imprints from 2009 and one of the 2013 "single" datasets are combined (e.g. C1 contains the imprints of *A* and B1). Further information on the concrete specifications can be found in [Ki16, KU17a, KU17b]. For all recognition experiments and datasets the same performance figures as in [KU17a] have been derived to evaluate the recognition results. For the evaluation process of the recognition accuracy, the Fingerprint Verification Contests' (FVC) procedure was performed, see [Ma09].
In the following, we describe the different techniques applied to separate (minutiae) data resulting from the currently acquired FP and the already present ghost FP.

[4] http://www.nist.gov/itl/iad/ig/nbis.cfm
[5] http://www.neurotechnology.com/verifinger.html
[6] http://biometrics.idealtest.org/dbDetailForUser.do?id=7

Masking the Background (MBw): This first method is used to separate the background and region of interest (ROI) of the FP images from each other by applying FP segmentation. After sharpening the edge information we used a Sobel operator to retrieve the edges of the ROI. We also tested other edge detection algorithms (e.g. Canny Edge detector, Prewitt operator and Harris corner points as used in [WTG07]), but for the given data, the Sobel approach worked best. Subsequently performing image dilation and erosion calculations we obtained the final masks. In Figures 1a) and b) an imprint mask and the combination of mask and image is displayed.

Smooth Masking of the Background (SMB): This approach was designed to enhance the background masking method (MBw). According to the fact that the edges of the masks could introduce new positions where minutiae information may be detected falsely, a Gaussian smoothing operation using $\sigma = 2$ as parameter was applied. In Figures 1c) and d) the example image of user 7 can be seen.

(a) Imprint mask. (b) Mask and image combined. (c) Imprint smooth mask. (d) Smooth mask and image combined.

Fig. 1: Background masked fingerprint images of user 7, dataset *B*4.

Splitting the ROI and Background minutiae (ROIm): This method was designed to perform a reference analysis for the background masking method in order to mitigate for newly created minutiae caused by the masking operation. For that reason we created the minutiae files, then we used the background masks to separate the minutiae which have been detected in the background and in the ROI. The selected minutiae were stored in two single files and we repeated the matching process using NBIS on the background and the ROI minutiae independently. Results are provided for the ROI minutiae only, as background minutiae do not lead to sensible recognition results.

Removing "stable" ROI and Background minutiae (wS and ROIwS): The previously introduced approaches are focusing on removing artifacts caused by ghost FPs by focusing on the ROI only - spatial background information is removed. However, ghost FP might also affect the ROI of course. To discriminate minutiae resulting from ghost FP from minutiae of the current imprint, we introduce the concept of "stable minutiae". While for taking different imprints of the same finger the finger is lifted off the sensor and re-allocated each time the data is acquired (causing the FP minutiae to manifest at different spatial locations), this is not the case for minutiae caused by ghost FPs, as these are detectable at the same x- and y- axis position (as long as the sensor is not cleaned minutiae information of some previous acquisition of the same finger remained on the sensor plate). According

to a visual analysis it could be confirmed that there is FP information of the same finger from a previous acquisition present in most of the cases (see Figure 2 as example). In the presented images minutiae in the ROI are coloured red and blue if they belong to the background. If a minutia is marked as stable it is coloured green (ROI) or magenta (background).

Fig. 2: Images with "stable" minutiae (first two images from the left) and ghost fingerprints.

FP recognition, using NBIS minutiae files without stable features (these are explicitly removed), was performed in two different ways. For the first case, we removed the stable minutiae information in the entire minutiae files. This led to results using all the minutiae detectable in the whole images, except the removed stable ones. We abbreviated this method with "wS" as acronym for "withoutStable". In the second approach we only focused on the ROI area for recognition and removed the stable minutiae there. The corresponding abbreviation is "ROIwS".

In Table 1 the number of images where stable minutiae information can be detected is presented in column *all images* (together with the relative amount of images in percent). In columns *all minutiae*, *ROIm* and *ROIwS* the average number of detected minutiae is displayed as well as the standard deviation concerning the minutiae appearance in the selected methods. In column *ROIm* the results considering only minutiae within the ROI exhibit a clear difference compared to using the whole imprints. According to the fact that ghost FPs are present in nearly all images of the datasets it is understandable why the mean values in *ROIm* are lower as in the *all minutiae* case. In terms of the standard deviation only minor fluctuations can be observed. The same minor variations can be detected in *ROIwS*. It seems that stable features are rarely in the imprints' ROI, which could be a disproof of the assumption that stable minutiae are responsible for effects exhibiting higher errors. Nevertheless, we considered this set-up in the recognition process as well because we wanted to prove/disprove the statement entirely.

Tab. 1: Number of images with "stable" minutiae and minutiae counts of all detectable minutiae.

dataset	*all images*	*all minutiae*		*ROIm*		*ROIwS*	
		μ	σ	μ	σ	μ	σ
A	504 (51.43%)	59.98	15.56	46.61	13.21	46.31	13.15
*B*1	180 (18.36%)	56.05	19.62	52.86	18.43	52.02	18.48
*B*2	364 (37.14%)	59.42	18.90	47.14	16.14	46.72	16.06
*B*3	500 (51.02%)	69.17	19.28	52.43	17.16	52.02	17.11
*B*4	416 (42.44%)	69.56	21.89	56.31	20.27	55.82	20.19
*B*5	246 (25.10%)	64.57	25.17	59.78	25.18	59.04	25.16

4 Experimental Evaluation

The most important results (the EER values for the different experimental set-ups) are presented in Table 2. In the first two columns the reference results, which have been calculated by analogy to [KU17a], are displayed. The differences in NEURO results as compared to the original ones of [KU17a] are caused by the usage of different SDK releases. The following columns represent the various experimental outcomes we obtained in this study. The best results are highlighted in bold numbers. The most obvious observation using NBIS is that the method MBw leads to a clearly worse EER performance compared to the reference values. This fact is not only valid for the single datasets, but also for the crossed ones in all cases. Further, the removal of ghost FPs does enhance the performance if it is done in a smooth way using some Gaussian filtering (SMB) because comparable measures can be reported for that case independently from NBIS and NEURO. Additionally, it is observable that the removal of stable features as it is done in wS and ROIwS experiments hardly influences the performance. According to that it can be concluded that the experiments we performed in removing ghost FPs did not have any impact on the higher error rates for time separated data in case of the EER. This performance figure is much higher for the time-separated datasets once more. But, the de-masking of ghost FPs does have an impact on the EER if it is done in a very rough way because new minutiae are introduced falsely (see MBw vs. SMB results). We also performed FP recognition using only the background information for all the described methods, but it was not possible to get EER values below (49%). Apart from that, it is interesting to observe that the usage of NEURO on dataset $C1$ indicates extraordinary cross-sensor effects, which have not been reported in [KU17a]. This must be caused by the different release we used. In the following we

Tab. 2: EER results of all datasets using NBIS and NEURO.

dataset	**entire images**		**MBw**		**SMB**		**ROIm**	**wS**	**ROIwS**
	NBIS	*NEURO*	*NBIS*	*NEURO*	*NBIS*	*NEURO*	*NBIS*	*NBIS*	*NBIS*
	single - all matching scores								
A	7.42	1.58	9.94	2.42	7.47	1.59	7.63	7.45	7.67
*B*1	8.95	2.77	10.98	2.84	9.71	2.58	8.98	8.93	9.09
*B*2	8.17	0.74	9.07	0.91	7.78	**0.66**	7.64	8.17	8.50
*B*3	9.07	3.06	11.68	3.34	8.99	3.03	9.40	9.24	9.35
*B*4	5.96	0.99	6.82	1.01	6.34	1.04	**5.70**	6.18	5.81
*B*5	7.30	1.29	9.65	1.61	7.82	1.42	7.59	8.23	7.53
	crossed - all matching scores								
*C*1	**12.63**	21.09	15.56	21.61	14.09	21.21	13.15	14.01	13.15
*C*2	14.76	4.55	17.79	5.02	14.99	4.42	14.43	14.85	14.46
*C*3	14.37	4.61	17.24	4.63	14.42	4.43	13.77	14.43	13.74
*C*4	13.18	**3.83**	15.66	4.10	13.35	3.93	12.94	13.26	12.97
*C*5	13.46	4.61	16.66	4.78	13.48	4.53	12.86	13.51	12.86

are going to discuss the other performance figures: Average Genuine Scores (AGS), Average Impostor Scores (AIS), the lowest FRR for FAR less or equal to 0.1% (FAR_{100}), and Zero False Acceptance Rate (ZeroFAR). The results can be looked up in Figure 3. At first we want to discuss the most important observation concerning a possible template ageing effect based on the AGS values: The decrease in the genuine scores is detectable for all performed NBIS and NEURO experiments independently. This is observable in Figures

3a) and b). There are fluctuations depending on the used dataset and analysis method, but the overall trend is similar. It is confirmed that NEURO exhibits some cross-sensor effects in dataset $C1$ because comparing images of the same finger involving the time-span leads to much lower genuine scores as can be seen by matching images of the same year. According to that the AGS for $C1$ is much lower compared to all the other datasets. For the average impostor scores (AIS) (see Figures 3c) and d)) a very similar stable behaviour as detected in [KU17a] can be described for the NBIS system. In case of NEURO there are some dataset dependent fluctuations which are based on the used datasets. In general it is interesting to observe that the crossed datasets' AIS is lower as in the single datasets from 2013. The experiments' FAR_{100} can be looked up in Figures 3e) and f). For both recognition methods it can be reported that the FAR_{100} is higher in all crossed datasets. Using NBIS the MB's performance figure is always worse compared to the others and some minor fluctuations can be detected for the other analysis methods. The high amount of variation is not describable in the NEURO case. Finally, we are having a look at the ZeroFAR values which are displayed in Figures 3g) and h). In general, the ZeroFAR for the crossed cases is always higher as for the single datasets. Nevertheless it must be mentioned that especially the results of $B3$-$B5$ and $C3$-$C5$ are much higher compared to the remaining values of the other datasets.

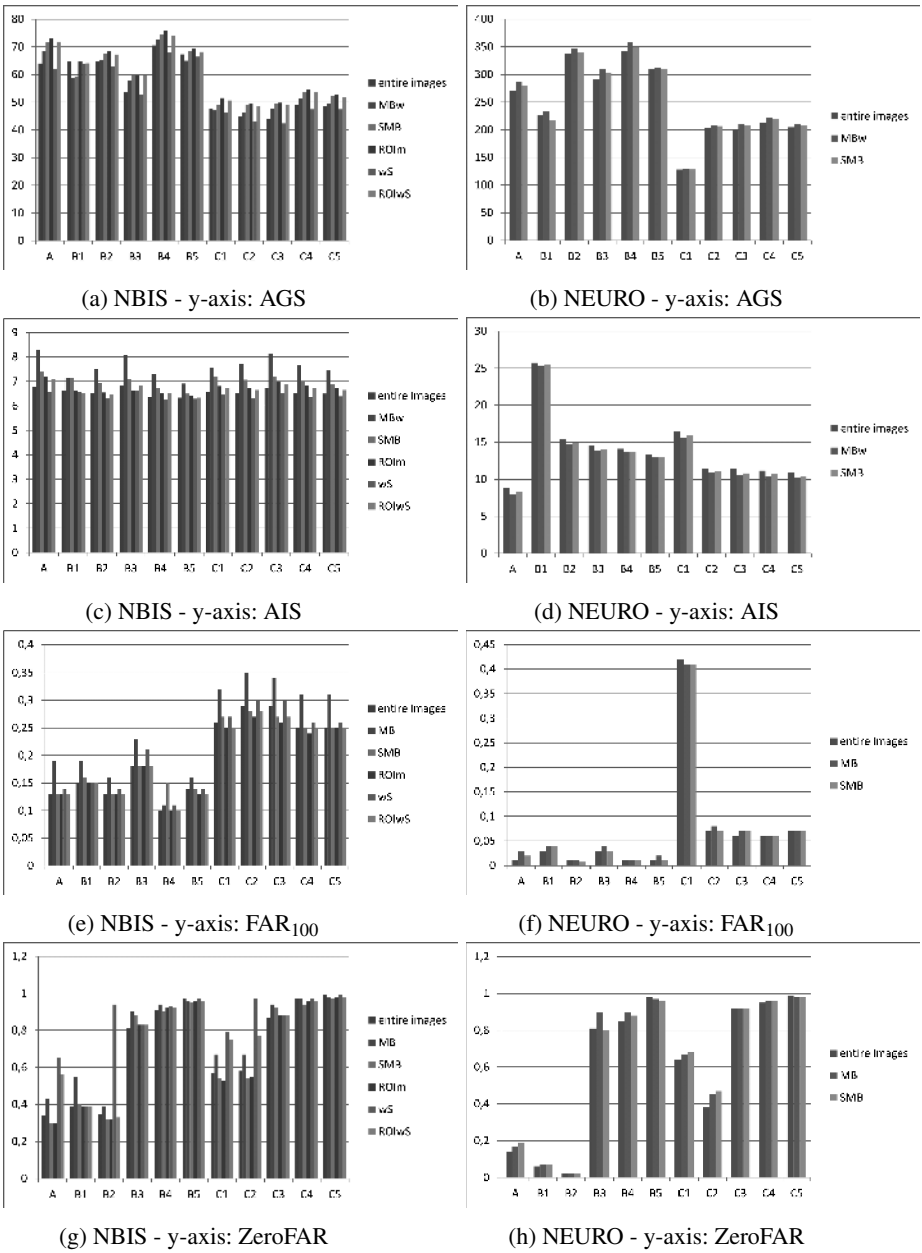
5 Conclusion

Based on the fact that in the given data a high number of ghost FPs (and thus stable minutiae) can be reported, it was a likely assumption that these might be responsible for the EER increase and average genuine score decrease in FP images exhibiting a time-span of 4 years. According to the knowledge that ghost FPs cause problems in FP segmentation (see [THG16, WTG07, Zh06]) and that double biometric identities influence the recognition process (see [FCM17]) the erroneous ghost FP information was removed using various methods. However, the same tendencies with respect to higher error rates, in particular increased EER and FRR caused by decreased genuine matching scores can be detected also with removed ghost FPs in our time-separated data. This leads to the disprove of the assumption that the observed effects are caused by ghost FP and corresponding stable minutiae information. This leads to the final statement that something different must cause the observed effects. So far it is not even clear, if decreased recognition accuracy as observed on the time separated data considered is caused by time-related or not time-related changes (i.e. differentiating between template ageing or a simple template change effect).

References

[ABI05] Arnold, M.; Busch, C.; Ihmor, H.: Investigating performance and impacts on fingerprint recognition systems. In: Information Assurance Workshop, 2005. IAW '05. Proceedings from the Sixth Annual IEEE SMC. pp. 1 – 7, june 2005.

[BG04] Bevilacqua, A.; Gherardi, A.: Age-related skin analysis by capacitance images. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. volume 2, pp. 703 – 706 Vol.2, aug. 2004.



(a) NBIS - y-axis: AGS (b) NEURO - y-axis: AGS

(c) NBIS - y-axis: AIS (d) NEURO - y-axis: AIS

(e) NBIS - y-axis: FAR_{100} (f) NEURO - y-axis: FAR_{100}

(g) NBIS - y-axis: ZeroFAR (h) NEURO - y-axis: ZeroFAR

Fig. 3: NBIS and NEURO performance figures of the experiments (x-axis: datasets).

[Do98] Doddington, George R.; Liggett, Walter; Martin, Alvin F.; Przybocki, Mark A.; Reynolds, Douglas A.: SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In: Int'l Conf. Spoken Language Processing (ICSLP). 1998.

[Dr12] Drahansky, M.; Dolezel, M.; Urbanek, J.; Brezinova, E.; Kim, T-H.: Influence of Skin Diseases on Fingerprint Recognition. Journal of Biomedicine and Biotechnology, 2012:Article ID 626148, 2012.

[FCM17] Ferrara, Matteo; Cappelli, Raffaele; Maltoni, Davide: On the Feasibility of Creating Double-Identity Fingerprints. IEEE Transactions on Information Forensics and Security, 12(4):892–900, 2017.

[Ga92] Galton, Francis: Finger Prints. Macmillan, London, 1892.

[GJ98] Gniadecka, M.; Jemec, G.: Quantitative evaluation of chronological ageing and photoageing in vivo: studies on skin echogenicity and thickness. British J. of Dermatology, 139:815–821, 1998.

[Go11] Gottschlich, C.; Hotz, T.; Lorenz, R.; Bernhardt, S.; Hantschel, M.; Munk, A.: Modeling the Growth of Fingerprints Improves Matching for Adolescents. Information Forensics and Security, IEEE Transactions on, 6(3):1165 –1169, sept. 2011.

[JRP99] Jain, A. K.; Ross, A.; Pankanti, S.: A prototype hand geometry-based verification system. In: Proceedings of the 2nd International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'99). pp. 166–171, 1999.

[KB09] Kekre, H.B.; Bharadi, V.A.: Ageing Adaptation for Multimodal Biometrics using Adaptive Feature Set Update Algorithm. In: IEEE International Advance Computing Conference. pp. 535 –540, march 2009.

[Ki16] Kirchgasser, Simon: Ageing Effects in Fingerprint Recognition. Master's thesis, Department of Computer Sciences, University of Salzburg, Austria, January 2016.

[KU16] Kirchgasser, Simon; Uhl, Andreas: Biometric Menagerie in Time-Span separated Fingerprint Data. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'16). Darmstadt, Germany, pp. 1–12, 2016.

[KU17a] Kirchgasser, Simon; Uhl, Andreas: Template Ageing and Quality Analysis in Time-Span separated Fingerprint Data. In: Proceedings of the IEEE International Conference on Identity, Security and Behavior Analysis (ISBA '17). New Delhi, Indien, pp. 1–8, 2017.

[KU17b] Kirchgasser, Simon; Uhl, Andreas: Template Ageing in Non-minutiae Fingerprint Recognition. In: Proceedings of the 5th International Workshop on Biometrics and Forensics (IWBF'17). Coventry, United Kindom, pp. 1–6, 2017.

[Ma05] Mansfield, AJ: , ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework, FDIS ed., JTC1/SC37/Working Group 5, Aug. 2005, 2005.

[Ma09] Maltoni, D.; Maio, D.; Jain, A.K.; Prabhakar, S.: Handbook of Fingerprint Recognition (2nd Edition). 2009.

[Ma12] Marcialis, Gian Luca; Didaci, Luca; Pisano, Alessandro; Granger, Eric; Roli, Fabio: Why template self-update should work in biometric authentication systems? In: Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on. IEEE, pp. 1086–1091, 2012.

[ME06] Modi, S.K.; Elliott, S.J.: Impact of image quality on performance: Comparison of young and elderly fingerprints. In: Proceedings of the 6th International Conference on Recent Advances in Soft Computing (RASC'06). pp. 449–454, 2006.

[Me13] Merkel, Ronny; Otte, Karen; Clausing, Robert; Dittmann, Jana; Vielhauer, Claus; Bräutigam, Anja: First investigation of latent fingerprints long-term aging using chromatic white light sensors. In: Proceedings of the first ACM workshop on Information Hiding and Multimedia Security. ACM, pp. 95–104, 2013.

[Mo07] Modi, S.K.; Elliott, S.J.; Whetsone, J.; Kim, H.: Impact of Age Groups on Fingerprint Recognition Performance. In: IEEE Workshop on Automatic Identification Advanced Technologies. pp. 19–23, 2007.

[PPP10] Popa, Gheorghe; Potorac, Romică; Preda, Nicolae: Method for fingerprints age determination. Romanian Journal of Legal Medicine, 18(2):149–154, 2010.

[RF05] Ribaric, S.; Fratric, I.: A biometric identification system based on eigenpalm and eigenfinger features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11):1698–1709, 2005.

[RJK07] Ryu, Jiuen; Jang, Jihyeon; Kim, Hale: Analysis of Effect of Fingerprint Sample Quality in Template Aging. In: NIST Biometric Quality Workshop II. nov 2007.

[SE05] Sickler, Nathan C; Elliott, Stephen J: An evaluation of fingerprint image quality across an elderly population vis-a-vis an 18-25 year old population. In: Security Technology, 2005. CCST'05. 39th Annual 2005 International Carnahan Conference on. IEEE, pp. 68–73, 2005.

[THG16] Thai, Duy Hoang; Huckemann, Stephan; Gottschlich, Carsten: Filter design and performance evaluation for fingerprint image segmentation. PloS one, 11(5):e0154160, 2016.

[Uc96] Uchida, Tateki; Komeda, Takashi; Miyagi, Masao; Koyama, Hiroyuki; Funakubo, Hiroyasu: Quantification of skin aging by three-dimensional measurement of skin surface contour. In: Systems, Man, and Cybernetics, 1996., IEEE International Conference on. volume 1. IEEE, pp. 450–455, 1996.

[UW09] Uhl, Andreas; Wild, Peter: Comparing Verification Performance of Kids and Adults for Fingerprint, Palmprint, Hand-geometry and Digitprint Biometrics. In: Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Application, and Systems 2009 (IEEE BTAS'09). IEEE Press, pp. 1–6, October 2009.

[UW13] Uhl, Andreas; Wild, Peter: Experimental Evidence of Ageing in Hand Biometrics. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'13). Darmstadt, Germany, September 2013.

[WF05] Woodard, D.L.; Flynn, P.J.: Finger surface as a biometric identifier. Computer Vision and Image Understanding, 100:357–384, 2005.

[WTG07] Wu, Chaohong; Tulyakov, Sergey; Govindaraju, Venu: Robust point-based feature fingerprint segmentation algorithm. Advances in Biometrics, pp. 1095–1103, 2007.

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

[Zh06] Zhu, En; Yin, Jianping; Hu, Chunfeng; Zhang, Guomin: A systematic method for fingerprint ridge orientation estimation and image segmentation. Pattern Recognition, 39(8):1452–1472, 2006.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fuŕ Informatik, Bonn 2017 95

Recognizing infants and toddlers over an on-production fingerprint database

Vanina Camacho [1], Guillermo Garella[2], Francesco Franzoni[3], Luis Di Martino [4], Guillermo Carbajal[5], Javier Preciozzi[6], Alicia Fernandez[7]

Abstract: It is widely known that biometric systems based on adults fingerprints have reached an outstanding performance when compared against other biometric traits. This explains their extensive use by governmental agencies in charge of citizen identification. Nevertheless, the performance is highly degraded when fingerprints of newborns or toddlers are used. In this work, we analyze the performance of existing solutions (both at sensor and matching level) using 45000 infants fingerprints taken from an on-production civilian database. We also propose a solution by zooming the input fingerprints with an interpolation factor based on ridges distances. The developed solution shows improvements in both fingerprint quality (NFIQ 2.0) as well as recognition performance.

Keywords: fingerprint, biometric, recognition, newborns, infants, toddlers, id, interpolation

1 Introduction

Fingerprints are commonly preferred over other biometric traits for their inherent features, such as *distinctiveness*, *permanence*, and *performance* [RFJ08]. This explains its extensive use by national IDs/passports issuance offices and borders control among others. Fingerprint matching solutions are very mature and achieve a very good performance. Nevertheless, most available systems and research focus on adult fingerprints. In recent years, several works were conducted to analyze the suitability of using fingerprints in children [Co13, Ja15, Ja16a, Ja16b]. To the best of our knowledge, the most extensive study was presented by the *Joint Research Center* of the *European Commission (UE)* [Co13]. In this work, a database of fingerprints obtained from $2,611$ children (in the 0-12 years old range) with 500 dpi scanners were used. This data was acquired by the Portuguese government passport issuance offices. The report concluded that it was difficult to identify children with less than six years old. They also concluded that it is necessary to use higher resolution scanners. Recently, a longitudinal study done in a population of 309 individuals ranging from zero to five years old was presented [Ja16b]. It was reported for the first time, the feasibility of using fingerprint to identify children at an early age: very good results

[1] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, vcamacho@fing.edu.uy
[2] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, guillermo.garella@fing.edu.uy
[3] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, francesco.franzoni@fing.edu.uy
[4] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, dimartino@fing.edu.uy
[5] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR carbajal@fing.edu.uy
[6] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR/ Dirección Nacional de Identificación Civil, jpreciozzi@dnic.gub.uy
[7] Instituto Ingeniería Eléctrica - Facultad de Ingeniería - UdelaR, alicia@fing.edu.uy

were obtained in children older than six months using special scanners of 1270 dpi. Good results were also obtained using standard adult fingerprint scanners of 500 dpi in children of at least one year old.
In this work, we conduct an extensive analysis using a considerable size dataset obtained from an on-production environment. The data was obtained from the Uruguayan National Identification Agency (DNIC[3]) which is responsible for issuing Uruguayan passports and ID cards. The main contributions of the present work are twofold. First, we continue the analysis done in previous studies [Co13, Ja15, Ja16a, Ja16b] but with a greater number of individuals. Secondly, we show the robustness of fingerprints to identify children on an on-production civilian database, where fingerprints were acquired during the standard ID card and Passport issuance processes.

The rest of the article is organized as follows. In Section 2 we describe the scenario in which the data was obtained. Section 3 describes the protocol used to perform the analysis. It includes a preprocessing proposal with a zoom stage, with a learned interpolation factor, based on the estimation of the ridge distance for an age range. Section 4 describes the experiment realized and results obtained. Finally, section 5 describes the conclusions and future work.

2 Description of the Uruguayan scenario

Uruguay has a population of 3.42 millions people, with a born rate of 14.15% [4]. The agency in charge of national identity management is the National Civil Identification Agency (DNIC) and is responsible of issuing ID cards and Passports. The identification process is based on fingerprint comparison, which, as usual, is used in two different scenarios: enrollment and renewal. Uruguay is a particular country regarding identification since the enrollment is done at birth: by law, parents have 45 days to obtain the identity card for the newborn. This is done since its creation in 1978. Due to the difficulty of matching the fingerprints of newborns, this information is only stored but not used for identity verification or de-identification. When the child is 5 years old, a complete ten fingerprint is obtained and these fingerprints are stored for this individual throughout his life. This procedure is in execution since 1978 which means that all the Uruguayans born before 1973, are actually enrolled with their 5 year old fingers (nearly 2.4 millions). It is worth noting that during the enrollment and the renewal process, an adult fingerprint may be compared with a fingerprint corresponding to a 5 year old child and this is done systematically and as part of the actual process. Until 2011, all the fingerprints were obtained using ink on paper. As part of the enrollment process, these templates were scanned at 500 dpi, segmented on each finger, and stored digitally to further visualization. It was not until 2010 that an AFIS system was installed on DNIC, which was filled with all these previously scanned fingerprint image. From 2011 until now, the ten fingerprints were acquired using fingerprint scanners. Given the fact that acquisition devices operate at one fixed resolution, there are certain disadvantages while working with children, mostly because of the small

[3] in Spanish: Dirección Nacional de Identificación Civil
[4] https://en.wikipedia.org/wiki/Uruguay

size fingerprint. Also, there are some problems with certain assumptions used by recognition algorithms with respect to average ridge distances. Another obstacle with smaller distances between fingerprint features is the decreased ability of algorithms to deal with the non-similarities introduced by distortion effects and bad positioning of the finger at capture.

3 Protocol Specification

Because of the national-wide characteristics of the biometric database, the number of individuals at our disposal is far more bigger than any other work done before.

3.1 Dataset

One of the criteria used to select the individuals and data for this study was obviously to have at least two fingerprints for the same finger. The total size of the dataset is 45000 pairs of fingerprints that were grouped by age as detailed on Table 1, where also the total number of individuals on each group is described.[5]. The variability on the number of individuals per group deserves an explanation. As explained before, it is mandatory for all children born in Uruguay to obtain an identity after 45 days of born. For this reason we have more than 13000 individuals on the group ranging from 0 to 1 month. In a typical scenario (if the child was registered in time and the document was not lost or stolen), the child has to renew the document at the age of 5. This is the reason we have more than 14000 individuals on the group between 5 and 6 years old. The other cases are exceptions on the typical identity management, and correspond to late enrollments or lost/stolen documents (it can also correspond to passport emission, which is also included in the same database). For these reasons, we have different number of fingerprints captures for each group. In this work, data is distributed according to the age at the time of the first capture. In table 1 we show the distribution of data and the number of fingerprints pairs for each set. All fingerprints were acquired by a well known commercial scanner model at 500 dpi, used by DNIC in all its offices (35 in total, distributed all along the country). Finally, and to compare some of the results with public databases, we also present some results using the adults NIST database MFCP2 [WF16].

3.2 Preprocessing method

One of the reasons why children identification is challenging is that most of the commercial systems are implemented and configured to work with adult fingerprints. This was already reported in [Ja16b] using NFIQ 2.0 as a quantitative measure of the quality of the fingerprints. A similar result was obtained in this work, as can be seen in Table 2. In order to use existing commercial AFIS systems, we need to preprocess the children fingerprints

[5] Due to privacy regulations, DNIC data cannot be published.

in such a way that the resulting image is well suited for these AFIS systems. This preprocessing process consists of two steps: an interpolation (to resize children fingerprints to an adult size) and segmentation (to reduce errors on minutiae extraction). Both steps are explained in the next sections.

Interpolation: In [Ja15, Ja16b], a fixed interpolation value of 1.8 was used in all cases, for fingerprints acquired at 500 dpi. In this work, we try to obtain a scale factor that depends on the age, and apply this obtained scale factor to resize the image to adult size. Even when we compare two children fingerprints, we rescale both of them to an adult size, enabling the use of existing AFIS systems. To determine the scale factor for each age, we follow [Kh11]: knowing the local ridge orientation, distance between ridges is measured by projecting the gray value levels along an orthogonal direction to the local ridge orientation and finding the minimum values. For each one of the ages, the median of the distances between ridges was selected, which was later compared with the distance between adult ridges on a 500 dpi image (9 pixels). The relation is given by Equation 1.

$$f_{oi} = \text{distance between ridges on adults} / \text{distance between ridges on age group}(dc_i) \quad (1)$$

Table 1 shows the result of the ridge distance analysis, divided by age group. As expected, median value augment as age increase. Table 1 also includes the final interpolation values for each age, which are the ones used to interpolate fingerprints in all the experiments done in this work.

Age group at first capture	Total of fingerprints	Average age at first capture	Average age at second capture	Interpolation Factor	Distance between ridges
Newborns	13050	18 d	8 m	1.52	5.92
1-2 m	5395	1m 12 d	8 m	1.63	5.52
2-3 m	902	2m 12 d	8 m	1.65	5.43
3-4 m	349	3m 13d	8 m	1.58	5.68
5-6 m	195	5m 13d	10 m	1.60	5.62
6-12 m	627	8m 27d	1 y 2 m	1.54	5.86
1-2 y	988	1y 6m	2 y 2 m	1.49	6.06
2-3 y	1164	2y 6m	3 y 7 m	1.47	6.14
5-6 y	14836	5y 6m	7 y 2 m	1.32	6.80
6-7 y	2234	6y 6m	7 y 6 m	1.29	6.96
7-8 y	1397	7y 6m	8 y 8 m	1.26	7.13
8-9 y	1471	8y 8m	9 y 10 m	1.24	7.27
9-10 y	2787	9y 9m	10 y 7m	1.22	7.40

Tab. 1: Description of database, interpolation factor and distance between ridges for each set.(d=days, m = months, y = years, newborns $<$ 1 m)

Segmentation: In order to eliminate the background acquisition noise on paper captures, fingerprints are segmented by looking for the first n aligned points whose values are black enough, assuming that they belong to the fingerprint, and cropping the image to this bounds. Values n and the threshold to assume that pixels are enough black were learned from FVC 2004 [Fi04].

4 Experiments and results

4.1 Performance Evaluation Metrics

In order to analyze the system performance, we use the the usual metrics: *True acceptance rate* (TAR) which is the percentage of times that the system correctly verifies a true claim of identity, and *False Acceptance Rate* (FAR), which is the probability that the system incorrectly matches the input pattern to a non-matching template in the database. *Receiver Operating Characteristics* (ROC) curve is plotted as TAR vs FAR at different thresholds (from 0 to 1) to indicate the verification performance.

4.2 Quality vs accuracy

Performance of biometric systems depends to a great extent on the quality of data. Therefore, quality indicators can be used as a way to compare the effectiveness of different preprocessing methods. In this work we used NFIQ 2.0[De], which delivers a number from 0 to 100 directly related with the performance prediction of the matcher evaluating a single fingerprint. In table 2 NFIQ 2.0 data quality is shown.

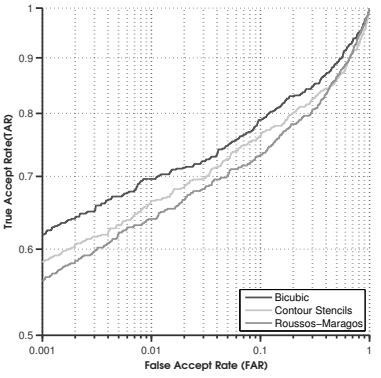
Age Group	Numbers of fingerprints	Initial Quality	Preprocessing quality	Variance of preprocessing quality
Newborns	2264	1,68	2,62	4,37
1-2 m	2176	2,25	6,98	9,19
2-3 m	733	2,66	9,83	11,55
3-4 m	288	3,46	8,17	11,95
5-6 m	161	3,59	10,37	9,97
6-12 m	482	6,12	14,16	15,07
1-2 y	712	14,50	30,23	20,53
2-3 y	784	23,55	42,83	22,96
5-6 y	2963	36,03	48,56	25,48
Adults (NIST MFCP 2)	1086	45,98	-	-

Tab. 2: NFIQ 2.0 data quality

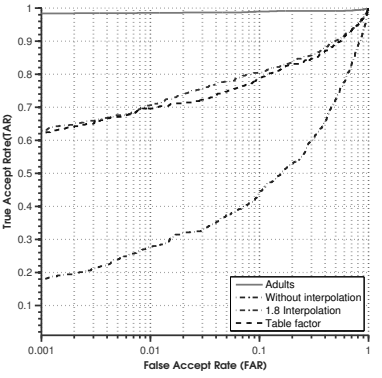
4.3 EXPERIMENTS

We start our set of experiments by analyzing interpolation. We compare a classic bi-cubic interpolation with two other interpolation methods: Interpolation with Geometric Contour Stencils [Ge11a] and Tensor-Driven Diffusion for Image Interpolation [Ge11b]. Figure 1(a) presents the results obtained on a one year old database with 720 fingerprints. In all cases, we use the interpolation factor described in Table 1. We can see that there is no significative difference between the different methods and in fact, bi-cubic obtains the better results. For the rest of the experiment, we use bi-cubic as the interpolation method.

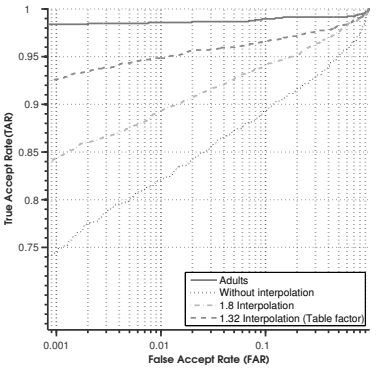
In the next experiment, we analyze the results on a database of one and five years old and compare them with the performance on an adult database (NIST MFCP2 database [WF16]). Table 3, Figures 1(b) and 1(c) presents the results. In this case, each pair of fingerprint is considered an identity (we present the results considering two fingerprints per identity later). It is clear from the results that interpolation is mandatory to obtain good results. What is more, applying the correct interpolation factor improves the results in the case of five years, as we can see when we compare the results obtained using the interpolation factor from Table 1 and the one obtained with an interpolation factor of 1.8. In the case of one year, we obtain almost the same performance for both interpolation factors. Since selected minutia extractor works with a default image size, using the proposed inteprolation factor we ensure looking for minutias over the whole fingerprint. In our final



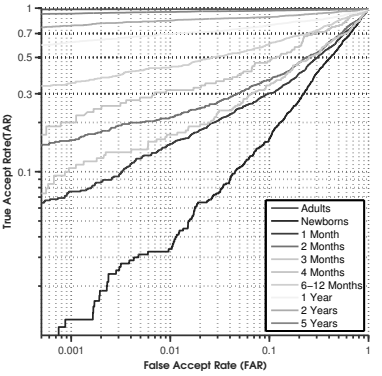
(a) Different interpolation performances over a year old base with 720 fingerprints



(b) Comparison between adults and 1 year old performance



(c) Comparison between adults and 5 year old performance



(d) Performance of every set interpolating by table I factors

Fig. 1

experiment, we compare the results obtained using the corresponding interpolation factor obtained from Table 1 for different databases grouped by age. Figure 1(d) and Table 4 present the results. We also include the results obtained for an adult database. From the

	IF=1	IF proposed	IF=1,8	Fusion IF proposed	Fusion IF=1.8
TAR (%) five years	74,41	92,64	84,35	98.33	90.65
TAR (%) one year	18.12	61,88	62.34	79.28	81.42
TAR(%) Adults	98.39	-	-	-	-

Tab. 3: Performance given by TAR for a fixed FAR in 0.1% for interpolation factor of 1.8 and the proposed, with and without fusion for two fingers, (IF = interpolation factor)

results, we can see that from 5-6 years old (92,64%) are comparable to the ones obtained for adults (98,39%) and even 1-2 years and 2-3 years present good results (61,88% and 78,37%). We recall that in all these experiments, we consider that each identity has only one fingerprint. In order to compare our results with the one obtained in [Ja16b], we perform a last experiment where we consider two fingerprints for each individual (right thumb and right index). In the five years old database, where we have 599 individuals, we obtain a TAR of 98.33% at a fixed FAR in 0.1%. From a total of 111 subjects in one year database, we obtain a TAR of 79.28% at a fixed FAR in 0.1%. In [Ja16b], authors reported a TAR of 100% for a fixed FAR in 0.1% for children from one to five years old. When we replicate the experiment with our dataset (applying 1.8 factor), we obtain a TAR 90.65% for the five years old and a TAR of 81.42% for one year old database, in both cases with a FAR of 0.1%. We believe that the main differences with the result reported in [Ja16b] is obviously the source of the dataset. In our case, the data was obtained directly from the on-production environment, without any participation on the way fingerprints were acquired. We consider that the results obtained from the fusion experiment (which is in fact the usual scenario on identification, where in general we have more than one fingerprint per individual) are very illustrative and confirms that fingerprints can be used to identify children starting from one year old. This claim is supported with the data used in this work, obtained directly from an on-production system.

Age Group	TAR (%)	TAR (%) with preprocessing
Newborns	NA	1,25
1-2 m	NA	7,57
2-3 m	NA	15,61
3-4 m	NA	10,53
5-6 m	NA	20,00
6-12 m	2,53	34,88
1-2 y	18,12	61,88
2-3 y	27,24	78,37
5-6 y (2000)	74,41	92,64
NIST MFCP 2	98,39	98,39

Tab. 4: Performance given by TAR for a fixed FAR in 0.1%

5 Conclusions and future work

In this work, we present an analysis of using fingerprints for children identification and verification. We perform all the study on a production database, where fingerprints were

acquired on usual ID card and Passport. The results show that fingerprints can be used without any additional hardware starting from one year old. As we can see in Table 4, performance improves in accordance with children's growth. We also show that applying the corresponding interpolation factor, we obtain similar or better results than using a fixed interpolation size. We conclude that preprocessing fingerprint according to their age is a necessary step that deserves more research.

In future works, we plan to determinate the system performance using the interpolation factor corresponding to each fingerprint ridges distance more than to a range according to age. We also want to acquire fingerprints with a scanner with a higher resolution in order to analyze the feasibility of using fingerprints for children below one year old. Because we have access to the full fingerprint database at DNIC, we are planning to repeat the experiments with far more individuals including matching between children and adults.

6 Acknowledgment

The authors would like to thank the DNIC agency for its collaboration and for granting us the permission to access their valuable data. This work has been supported by an investigation grant provided by the ANII (Uruguay Agency of Investigation and Innovation).

References

[Co13] Commission, European: Fingerprint Recognition for Children. Technical report, Joint Research Centre, 2013. Available: https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/fingerprint-recognition-children.

[De] Development of NFIQ 2.0. Available: https://www.nist.gov/services-resources/software/development-nfiq-20.

[Fi04] Fingerprint Verification Competition (FVC), 2004. Available: http://bias.csr.unibo.it/fvc2004/.

[Ge11a] Getreuer, Pascal: Image Interpolation with Geometric Contour Stencils. Image Processing On Line, 1:98–116, 2011.

[Ge11b] Getreuer, Pascal: Roussos-Maragos Tensor-Driven Diffusion for Image Interpolation. Image Processing On Line, 1:178–186, 2011.

[Ja15] Jain, Anil K; Arora, Sunpreet S; Best-Rowden, Lacey; Cao, Kai; Sudhish, Prem Sewak; Bhatnagar, Anjoo: Biometrics for Child Vaccination and Welfare: Persistence of Fingerprint Recognition for Infants and Toddlers. arXiv preprint arXiv:1504.04651, 2015.

[Ja16a] Jain, Anil K.; Arora, Sunpreet S.; Best-Rowden, Lacey; Cao, Kai; Sudhish, Prem S.; Bhatnagar, Anjoo; Koda, Yoshinori: Giving Infants an Identity: Fingerprint Sensing and Recognition. In: Proceedings of the Eighth International Conference on Information and Communication Technologies and Development. ICTD '16, ACM, New York, NY, USA, pp. 29:1–29:4, 2016.

[Ja16b] Jain, Anil K; Arora, Sunpreet S; Cao, Kai; Best-Rowden, Lacey; Bhatnagar, Anjoo: Fingerprint Recognition of Young Children. IEEE Transactions on Information Forensics and Security, 2016.

[Kh11] Khan, Muhammad Ali: , Fingerprint image enhancement and minutiae extraction, 2011.

[RFJ08] Ross, Arun A.; Flynn, Patrick J.; Jain, Anil K.: Handbook of Biometrics. Springer, 2008.

[WF16] Watson, Craig; Flanagan, Patricia: NIST Special Database 14 Mated Fingerprint Card Pairs 2 WSQ Compressed Images. 2016.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 105

Benchmarking Fingerprint Minutiae Extractors

Tarang Chugh[1], Sunpreet S. Arora[2], Anil K. Jain[1], Nicholas G. Paulter Jr.[3]

Abstract: The performance of a fingerprint recognition system hinges on the errors introduced in each of its modules: image acquisition, preprocessing, feature extraction, and matching. One of the most critical and fundamental steps in fingerprint recognition is robust and accurate minutiae extraction. Hence we conduct a repeatable and controlled evaluation of one open-source and three commercial-off-the-shelf (COTS) minutiae extractors in terms of their performance in minutiae detection and localization. We also evaluate their robustness against controlled levels of image degradations introduced in the fingerprint images. Experiments were conducted on (i) a total of 3,458 fingerprint images from five public-domain databases, and (ii) 40,000 synthetically generated fingerprint images. The contributions of this study include: (i) a benchmark for minutiae extractors and minutiae interoperability, and (ii) robustness of minutiae extractors against image degradations.

Keywords: fingerprint recognition, minutiae extraction, robustness to noise, interoperability

1 Introduction

A fingerprint recognition system typically comprises of four major modules: image acquisition, preprocessing, feature extraction, and matching (See Fig. 1). The errors introduced in each of these four modules, from image acquisition to matching cumulatively impact the overall system recognition performance. For instance, the low fidelity[4] of a fingerprint signal acquired by a sensor can introduce errors in preprocessing, induce poor feature extraction, and ultimately deteriorate the matching performance. Therefore, it is important to perform a comprehensive evaluation of each module independently to improve the overall performance of the fingerprint recognition system.

Fingerprint sensor certification standards (*e.g. PIV-071006* [Ni06] and *Appendix F* [Ni05]) mandate independent evaluation of fingerprint sensors. Hence vendors are required to demonstrate that their sensors can acquire a high-fidelity image with low-noise characteristics. Existing studies have evaluated the performance of sensors in terms of their resilience to external environmental factors (temperature and humidity), intrinsic subject-dependent factors (skin humidity and pressure) [Ka03], operational quality [CFM08], their interoperability [Al08], and finger liveness detection [Gh13]. Arora et. al [Ar16] have designed

This research was supported by grant no. 60NANB11D155 from the NIST Measurement Science program.

[1] Tarang Chugh and Anil K. Jain are affiliated with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824. Email: {chughtar, jain}@cse.msu.edu

[2] Sunpreet S. Arora is currently affiliated with the Emerging Technology, Risk and Authentication Products Group, Visa Inc., Foster City, CA, 94404. Email: sunarora@visa.com. At the time this research was conducted, Sunpreet was affiliated with the Dept. of Computer Science and Engineering, Michigan State University.

[3] Nicholas G. Paulter Jr. is affiliated with the National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD, 20899. Email: paulter@nist.gov

[4] Fidelity refers to the degree of exactness with which friction ridge patterns on a finger are reproduced by the sensor

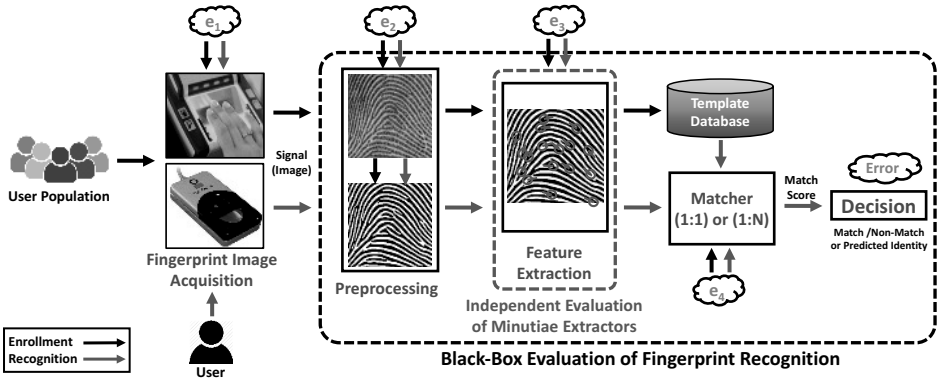


Figure 1: Framework of a typical fingerprint recognition system. While existing studies evaluate the recognition system from an end-to-end perspective, we provide a benchmark for minutiae extraction module. Errors introduced at different steps of the system, *i.e.* fingerprint acquisition (e_1), preprocessing (e_2), minutiae extraction (e_3), and matching (e_4), cumulatively impact the overall performance.

and fabricated 3D fingerprint targets and whole hand targets for repeatable evaluation and calibration of fingerprint sensors.

On the contrary, studies pertaining to fingerprint preprocessing, feature extraction, and matching, evaluate these modules in entirety as a black-box with the goal to improve the overall matching performance. National Institute of Standards and Technology (NIST) conducts fingerprint vendor technology evaluations (FpVTE) to benchmark the capabilities of fingerprint recognition systems in terms of identification accuracy and computational requirements [Wi04, Wa14]. The 2014 FpVTE [Wa14] reports that the best performing system achieved a FNIR of 1.9% for single index finger, and 0.09% using all ten-fingers, at a FPIR of 0.1%. Fingerprint verification competitions[5] (FVC 2000-2006) also evaluate systems from an end-to-end perspective. Although these third-party evaluations are useful, they do not evaluate individual modules. For instance, in the case of a false match or a non-match, it is uncertain whether the error is caused due to poor image quality, minutiae extraction errors, or inability of the matcher to handle distortion. An independent evaluation of the individual modules will enable us to understand the error sources and design an interoperable system.

It is generally known that minutiae extraction is critical to fingerprint recognition accuracy. Minutiae-based representation is the most widely used approach, essentially due to its (i) interpretability, (ii) high matching performance, (iii) storage efficiency, (iv) applicability to match fingerprints/latents in forensic casework, and (v) evidential value (*i.e.* expert testimony based on mated minutiae is admissible in the courts of law) [JFN10]. The FVC-onGoing [Do09], in addition to benchmarking performance at the system level, also provides benchmarks for (i) fingerprint orientation extraction, and (ii) matching standard minutiae-based templates [ISO/IEC 19794-2 (2005)]. However, accuracy and robustness evaluation of minutiae extracted using different minutiae extractors are needed in order to benchmark their performance and minutiae interoperability.

[5] https://biolab.csr.unibo.it/FVCOnGoing/UI/Form/Home.aspx

Minutiae interoperability tests (*e.g.* MINEX III [Fl15]) evaluate the compliance between minutiae-based template generators and matchers from different vendors. Kayaoglu et al. [KTU13] compared the matching performance based on automatically extracted minutiae and manually labelled minutiae. However, these tests did not evaluate the underlying factors limiting the minutiae interoperability, *i.e.* variations in the minutiae detection and localization ability. Moreover, the images input to minutiae extractors may contain distortion and motion blur due to variance in pressure applied on the sensor platen, and may have poor contrast due to dry/wet fingers (See Fig. 2). To address these challenges, this study conducts:

- A repeatable and controlled evaluation of minutiae extraction in terms of their detection and localization performance, for one open-source and three commercial minutiae extractors.
- A rigorous assessment of robustness of minutiae extractors in the presence of controlled levels of noise and motion blur to understand their limitations.

Figure 2: Challenges in automated fingerprint processing. Five different impressions of the same finger (from FVC2004 DB1A). These illustrate (a) reference fingerprint, (b) large non-linear distortion (compare the triangle in (b) to triangle in the reference fingerprint (a)), (c) smudged areas due to wet fingerprint, (d) and (e) broken ridge structure due to dry and noisy fingerprints.

2 Evaluation Protocol

2.1 Databases

The fingerprint images used in this evaluation study are grouped into two sets.

- Dataset-A contains 3,458 real fingerprint images compiled from five public domain databases: FVC 2002 (DB1A and DB3A), FVC 2004 (DB1A and DB3A) and NIST SD27 rolled prints database[6]. Each FVC database contains 800 fingerprint images (100 unique subjects, 8 acquisitions/subject), with ground truth minutiae marked by human subjects [KTU13]. NIST SD27 [NI] contains 258 rolled prints with ground truth minutiae marked by at least two certified forensic examiners.
- Dataset-B contains 40,000 synthetic fingerprints (including 5,000 unique masterprints, and 35,000 fingerprints degraded with controlled levels of noise and motion blur) generated using Novetta's biosynthetic software [No14]. It contains four levels of noise (including anatomical deformations, dryness, ridge noise) and three levels of motion blur.

Figure 3: Examples of fingerprint images from the six databases used in this evaluation study.

Figure 3 presents example fingerprint images from each of these databases. The two sets of fingerprint databases used in this study are summarized in Table 1. The average NIST Fingerprint Image Quality 2.0 (NFIQ 2.0) [Na16a], which lies in the range $[0,100]$ where 0 indicates the worst quality, and 100 refers to the best quality, is also presented for each database.

Database	**(# Fingerprints, # Subjects)**	**Ground Truth**	**Image Capture**	**Image Size ($h \times w$)**	**Avg. NFIQ2 value (s.d.)**
Dataset-A					
FVC2002 DB1A [Ma02]	(800, 100)	Manually Marked Minutiae	Optical sensor	374×388	64 (15)
FVC2002 DB3A [Ma02]	(800, 100)		Capacitive sensor	300×300	26 (13)
FVC2004 DB1A [Ma04]	(800, 100)		Optical sensor	480×640	59 (17)
FVC2004 DB3A [Ma04]	(800, 100)		Thermal sweep sensor	480×300	47 (16)
NIST SD27 (rolled prints) [NI]	(258, 258)		Digitized ink and paper	768×800	42 (10)
Dataset-B					
Synthetic masterprints [No14]	(5,000, 5,000)	N/A	Synthetically generated	480×512	71 (6)
Noisy prints [No14]	(20,000, 5,000)	Minutiae extracted from master prints	Synthetically generated	480×512	40 (23)
Motion blurred prints	(15,000, 5,000)		Synthetically generated	480×512	44 (26)

Table 1: A summary of fingerprint databases used in this evaluation study.

2.2 Evaluating Minutiae Detection and Localization

An ideal fingerprint minutiae extractor is expected to exhibit high precision in minutiae detection and localization, and minimize spurious and missing minutiae. We evaluate the performance of one open-source minutiae extractor *mindtct* [Na16b], and three minutiae extractors (COTS - A, B, and C) by comparing the extracted minutiae with the ground truth obtained from human subjects for Dataset-A. The performance of a fingerprint minutiae extractor depends heavily on the quality of input fingerprint images. Considering the large variations in the NFIQ 2.0 values, we segregate the fingerprint images from Dataset-A into five quality bins $[0,20]$, $[21,40]$, $[41,60]$, $[61,80]$, and $[81,100]$ based on the NFIQ 2.0 values. Figure 4 presents examples of fingerprint images corresponding to each of the 5 quality bins. For a fair evaluation, performance comparison between minutiae extractors is

[6] NIST SD27 is no longer publicly available.

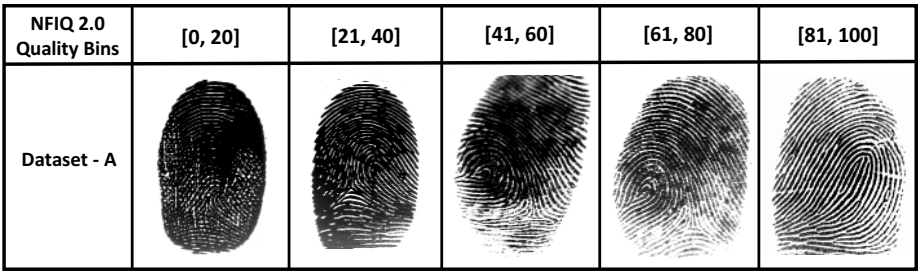


Figure 4: Examples of fingerprint images from Dataset-A corresponding to the 5 quality bins based on NFIQ 2.0 values, where $[0,20]$ represents the worst quality bin and $[81,100]$ indicates the best quality bin.

done only for fingerprint images within each quality bin. We do not utilize the synthetic fingerprint images (Dataset-B) for this evaluation, as the synthesis process itself introduces some spurious minutiae.

2.2.1 Minutiae Detection

Given a fingerprint image, let $F_d = \{f_d^1, f_d^2, ..., f_d^N\}$ be the set of N minutiae detected by a minutiae extractor, and $F_g = \{f_g^1, f_g^2, ..., f_g^M\}$ be the set of M ground truth minutiae marked by human subjects. A detected minutia f_d, and a ground truth minutia f_g are said to be *paired*, if f_d lies within a distance threshold δ around f_g. As the average ridge width for a 500 ppi fingerprint image is known to be approximately 9 pixels [Ma09], we fix the threshold to 10 pixels. If there is more than one detected minutia within the threshold, the one closest to the ground truth minutia is paired with it. In case of a tie, the pairing decision is made in favor of the minutia with smaller orientation difference. If a minutia has to be inserted in the set F_d, in order to pair it with a minutia in the set F_g, it is considered as a *missing* minutia. Similarly, if a minutiae in the detected set F_d, cannot be paired with any minutia in ground truth set F_g, it is deemed to be a *spurious* minutia. We utilize the Goodness Index (GI) metric of Ratha et al. [RCJ95] to evaluate the minutiae detection performance.

$$GI = \frac{\sum_{i=1}^{L} Q_i[P_i - D_i - I_i]}{\sum_{i=1}^{L} Q_i M_i} \tag{1}$$

where L = no. of 16×16 non-overlapping patches in the input image, Q_i = quality of the i^{th} patch ($good = 4,\ medium = 2,\ poor = 1$), P_i = no. of paired minutiae in the i^{th} patch, D_i = no. of spurious minutiae in the i^{th} patch, $D_i \leq 2 \cdot M_i$, I_i = no. of missing minutiae in the i^{th} patch, and M_i = no. of ground truth minutiae in the i^{th} patch, $M_i > 0$. In order to restrict the negative impact of outlier patches, the number of spurious minutiae (D_i) in a patch is restricted to a maximum value of $2 \cdot M_i$.

The quality index proposed by Chen et al. [CDJ05] is utilized. We do not consider patches with zero minutiae (near image boundary). The maximum value of GI is $+1$, which is obtained when $D_i = I_i = 0$ and $P_i = M_i$, *i.e.* all detected minutiae are paired and no. of detected and ground truth minutiae is the same. The minimum value of GI is -3, which is obtained when $P_i = 0$, $D_i = 2 \times M_i$, and $I_i = M_i$, *i.e.* no detected minutiae could be paired

and the no. of spurious minutiae takes its maximum possible value of $2 \cdot M_i$. Larger the value of Goodness Index, better the performance of a minutiae extractor. In addition to Goodness Index (GI), we also report the average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae.

2.2.2 Minutiae Localization

For a given minutiae extractor, let $\hat{f}_d = \{\hat{f}_d^1, \hat{f}_d^2, ..., \hat{f}_d^P\}$, $\hat{f}_d \subseteq F_d$, be a set of P detected minutiae points, paired with a subset of known ground truth minutiae points $\hat{f}_g \subseteq F_g$. The positional error (e_p) for the paired minutiae set $(\hat{f}_g, \hat{f}_d)$ is computed using the Root Mean Square Deviation (RMSD) [Tu11] given by:

$$e_p(\hat{f}_g, \hat{f}_d) = \sqrt{\frac{\sum_{i=1}^{P}[(x_g^i - x_d^i)^2 + (y_g^i - y_d^i)^2]}{P}} \tag{2}$$

where, (x_d^i, y_d^i) and (x_g^i, y_g^i) represent the locations of the detected minutia and the ground truth minutia, respectively. Similarly, the orientation error (e_θ) between the set of paired minutiae $(\hat{f}_g, \hat{f}_d)$ is computed using:

$$e_\theta(\hat{f}_g, \hat{f}_d) = \sqrt{\frac{\sum_{i=1}^{P}\phi(\theta_g^i, \theta_d^i)^2}{P}} \tag{3}$$

where $$\phi(\theta_1, \theta_2) = \begin{cases} \theta_1 - \theta_2 & \text{if} \quad -\pi \leq \theta_1 - \theta_2 < \pi \\ 2\pi + \theta_1 - \theta_2 & \text{if} \quad \theta_1 - \theta_2 < -\pi \\ -2\pi + \theta_1 - \theta_2 & \text{if} \quad \theta_1 - \theta_2 \geq \pi \end{cases}$$

2.3 Evaluating Robustness of Minutiae Extractors

The primary reason of errors in minutiae detection is the presence of artifacts due to variations in finger placement on the sensor platen, noise, finger moisture, fingerprint alterations, etc. A common evaluation technique, known as *stress testing*, is used to test a system beyond normal operating conditions, often to a breaking point. We evaluate the robustness of one open-source minutiae extractor *mindtct* [Na16b], and three commercial minutiae extractors in the presence of controlled levels of noise, finger dryness, and motion blur, to understand the stable operational conditions. We utilize the synthetic fingerprint images from Dataset-B for this evaluation.

2.3.1 Robustness against Noise

Fingerprint images acquired by the fingerprint readers may possess noise due to physical factors such as anatomical deformations in the friction ridge skin (scars, holes, scratches, etc.), finger moisture, and/or environmental contamination. These noise sources induce significant variation in minutiae extraction, even within multiple acquisitions of the same finger. To quantify the impact of noise on minutiae extractors, synthetic prints with controlled levels of noise are generated from synthetic master fingerprints. The noise model

in Novetta's biosynthetic software [No14] is utilized to add (i) anatomical deformations (scars, holes, and pressure variations), (ii) ridge noise (Perlin noise), and (iii) finger dryness. Fig. 5 presents different levels of noise added to a master fingerprint (used as the reference).

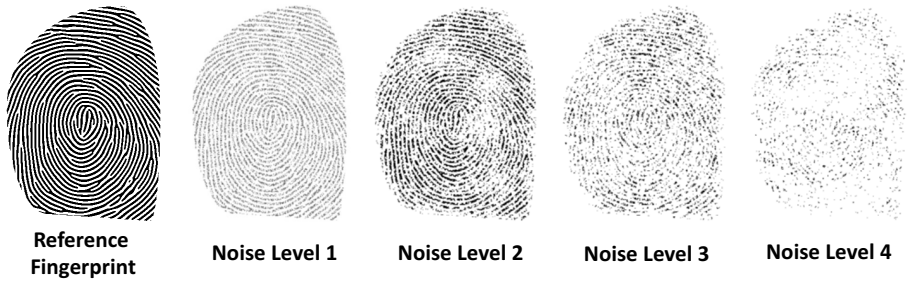


Figure 5: Four different levels of noise added to the master fingerprint (reference fingerprint).

2.3.2 Robustness against Motion Blur

Movements of the hand during fingerprint acquisition may lead to introduction of motion blur in the acquired image. We simulate three levels of motion blur in the synthetic master fingerprints by applying motion lens filter function in both horizontal and vertical direction [Li16]. The MATLAB functions $fspecial('motion', k)$ and $fspecial('motion', k, 90)$, with three different values of $k \in \{5, 7, \text{and } 9\}$ corresponding to increasing degrees of motion blur, are applied. Fig. 6 presents a synthetic master print and corresponding three different levels of motion blur.

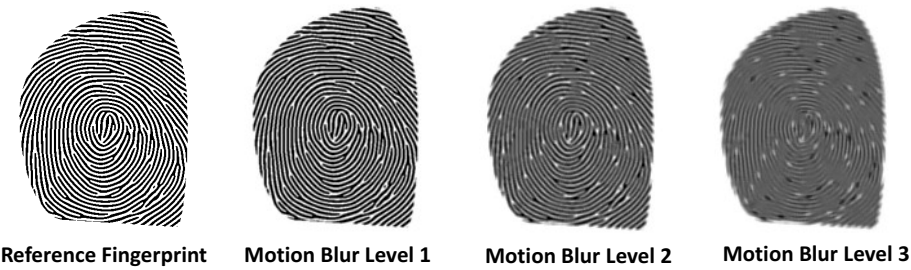


Figure 6: Three different levels of motion blur added to the master fingerprint (reference fingerprint).

3 Experimental Results

Goodness index, average positional error (e_p), and average orientation error (e_θ) are computed by comparing the output from one open-source minutiae extractor, *mindtct*, and three COTS minutiae extractors with the manually marked minutiae for Dataset-A, and minutiae extracted on the master print (without any image degradations) for Dataset-B.

3.1 Minutiae Detection and Localization

Fig. 7 presents an example fingerprint from FVC2002 DB1A dataset with overlaid manually marked minutiae and the extracted minutiae from one open-source minutiae extractor,

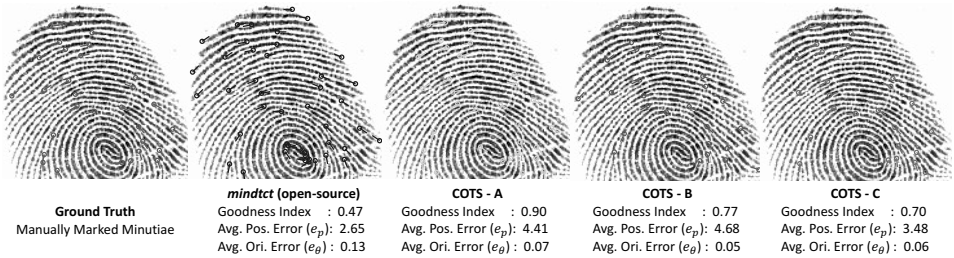


Figure 7: Example fingerprint from FVC2002 DB1A dataset with overlaid manually marked minutiae and minutiae extracted by four minutiae extractors (*mindtct*, and COTS A, B, and C). Goodness Index (GI) is unit less, while Avg. Positional Error (e_p) and Avg. Orientation Error (e_θ) are measured in pixels and radians, respectively.

NFIQ 2.0 Quality Bins	**Minutiae Extractor**	**[0, 20]**	**[21, 40]**	**[41, 60]**	**[61, 80]**	**[81, 100]**
# Fingerprints		419	803	1,051	1,053	132
Goodness Index Avg. (s.d.)	*mindtct*	−0.64 (0.77)	−0.45 (0.70)	−0.33 (0.59)	0.11 (0.38)	0.36 (0.25)
	COTS-A	−0.74 (0.69)	−0.14 (0.71)	0.00 (0.67)	0.47 (0.26)	0.60 (0.16)
	COTS-B	−0.03 (0.63)	0.22 (0.44)	0.33 (0.30)	0.48 (0.22)	0.57 (0.17)
	COTS-C	−0.04 (0.70)	0.12 (0.51)	0.21 (0.35)	0.40 (0.21)	0.48 (0.19)
Positional Error (e_p) **(in pixels)** Avg. (s.d.)	*mindtct*	3.95 (0.80)	3.78 (0.69)	3.60 (0.73)	3.22 (0.56)	3.10 (0.46)
	COTS-A	4.87 (0.66)	4.64 (0.61)	4.37 (0.64)	4.27 (0.60)	4.22 (0.59)
	COTS-B	4.53 (0.83)	4.24 (0.72)	4.02 (0.73)	4.00 (0.61)	3.89 (0.54)
	COTS-C	4.10 (0.86)	4.21 (0.82)	4.23 (0.78)	3.83 (0.70)	3.59 (0.57)
Avg. Orientation Error (e_θ) **(in rad.)** Avg. (s.d.)	*mindtct*	0.27 (0.23)	0.20 (0.12)	0.18 (0.09)	0.15 (0.06)	0.14 (0.04)
	COTS-A	0.16 (0.12)	0.13 (0.07)	0.12 (0.06)	0.11 (0.04)	0.10 (0.03)
	COTS-B	0.13 (0.13)	0.10 (0.06)	0.10 (0.05)	0.10 (0.04)	0.09 (0.03)
	COTS-C	0.14 (0.12)	0.11 (0.07)	0.10 (0.05)	0.10 (0.04)	0.09 (0.02)

Table 2: Performance comparison of four minutiae extractors (*mindtct*, and COTS A, B, and C) in terms of minutiae detection and localization accuracies. This evaluation utilizes fingerprint images (Dataset-A) from five public domain datasets, available with manually marked ground truth minutiae. Minutiae detection is measured in terms of Goodness Index (GI), a unit less measure in the range [-3, 1]. A large value of GI suggests high number of detected minutiae are paired with ground truth minutiae and low number of spurious or/and missing minutiae.

mindtct, and three COTS minutiae extractors. The values for the three performance metrics, Goodness Index, Positional Error, and Orientation Error are also reported for each minutiae extractor output. Tab. 2 presents a summary of the performance comparison between the four minutiae extractors in terms of minutiae detection and localization accuracies for Dataset-A. In comparison to other minutiae extractors, COTS-B consistently achieves a higher value of Goodness Index across all quality levels. Performance of COTS-A is observed to be highly dependent on fingerprint quality, as it achieves the lowest Goodness Index for low quality images (NFIQ 2.0 = [0, 20]), and highest Goodness Index for high quality images (NFIQ 2.0 = [81,100]). The open-source minutiae extractor, *mindtct*, achieves low Goodness Index compared to COTS minutiae extractors across all quality values, however, it also achieves lowest positional errors suggesting high positional accuracy for the paired minutiae. In general, a NFIQ 2.0 quality value lower than 20 leads to a negative Goodness Index and higher localization errors with larger variances. It can be observed that as the quality level increases, the Goodness Index values also increase, indicating higher

NFIQ 2.0 Quality Bins	Minutiae Extractor	[0, 20]	[21, 40]	[41, 60]	[61, 80]	[81, 100]
# Fingerprints		419	803	1,051	1,053	132
Paired Minutiae / Ground Truth (P_i / M_i) Avg. (s.d.)	*mindtct*	0.77 (0.12)	0.81 (0.11)	0.82 (0.09)	0.84 (0.08)	0.86 (0.07)
	COTS-A	0.77 (0.14)	0.79 (0.16)	0.78 (0.17)	0.85 (0.07)	0.86 (0.06)
	COTS-B	0.71 (0.15)	0.76 (0.12)	0.79 (0.10)	0.82 (0.08)	0.84 (0.07)
	COTS-C	0.74 (0.14)	0.74 (0.11)	0.75 (0.09)	0.77 (0.08)	0.78 (0.09)
Spurious Minutiae / Ground Truth (D_i / M_i) Avg. (s.d.)	*mindtct*	1.19 (0.63)	1.06 (0.60)	0.97 (0.53)	0.57 (0.34)	0.36 (0.21)
	COTS-A	1.29 (0.60)	0.72 (0.52)	0.56 (0.44)	0.22 (0.20)	0.12 (0.09)
	COTS-B	0.44 (0.45)	0.30 (0.31)	0.25 (0.21)	0.15 (0.13)	0.10 (0.08)
	COTS-C	0.52 (0.55)	0.36 (0.39)	0.30 (0.28)	0.13 (0.12)	0.09 (0.08)
Missing Minutiae / Ground Truth (I_i / M_i) Avg. (s.d.)	*mindtct*	0.23 (0.12)	0.19 (0.11)	0.18 (0.09)	0.16 (0.08)	0.14 (0.07)
	COTS-A	0.23 (0.14)	0.21 (0.16)	0.22 (0.17)	0.15 (0.07)	0.14 (0.06)
	COTS-B	0.29 (0.15)	0.24 (0.12)	0.21 (0.10)	0.18 (0.08)	0.16 (0.07)
	COTS-C	0.26 (0.14)	0.26 (0.11)	0.25 (0.09)	0.23 (0.08)	0.22 (0.09)

Table 3: Performance comparison of the four minutiae extractors (*mindtct*, and COTS A, B, and C) in terms of average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae for fingerprint images of different quality (Dataset-A).

number of paired minutiae and lower number of spurious and/or missing minutiae. Tab. 3 presents the performance comparison of the four minutiae extractors in terms of average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae. It can be observed that the open-source minutiae extractor produces a much higher percentage of spurious minutiae, but a much lower percentage of missing minutiae, compared to other COTS minutiae extractors.

3.2 Robustness against Image Degradations

Tab. 4 summarizes the performance comparison between the four minutiae extractors on robustness against different levels of image noise for Dataset-B. It can be observed that as the noise level increases, the Goodness Index decreases, and the avg. positional error and the avg. orientation error increases. In comparison to other minutiae extractors, COTS-A achieves a much higher Goodness Index, and low positional and orientation errors even in the presence of higher levels of image noise. All the minutiae extractors exhibit similar avg. positional errors, but a much higher variance is observed in the case of COTS-C. Tab. 5 presents the performance comparison between the four minutiae extractors in terms of average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae for images with different levels of noise. It can be observed that COTS-A achieved a very high percentage of paired minutiae and much lower percentage of missing minutiae, resulting in a high Goodness Index. In terms of spurious minutiae, *mindtct* is observed to consistently perform poorly across all noise levels compared to the COTS minutiae extractors, producing much higher percentage of spurious minutiae.

The performance comparison of the four minutiae extractors in terms of minutiae detection and localization accuracies for images degraded with different levels of motion blur is presented in Tab. 6. It is observed that COTS-A achieves high Goodness Index value compared to other minutiae extractors with low avg. positional and orientation errors. In general, higher level of motion blur results in large negative values of Goodness Index for

Noise Levels	Minutiae Extractor	Level 1	Level 2	Level 3	Level 4
Goodness Index Avg. (s.d.)	*mindtct*	0.36 (0.27)	0.09 (0.32)	−0.43 (0.33)	−0.80 (0.25)
	COTS-A	0.80 (0.12)	0.72 (0.14)	0.52 (0.21)	0.15 (0.37)
	COTS-B	0.53 (0.19)	0.43 (0.21)	0.19 (0.23)	−0.15 (0.30)
	COTS-C	0.72 (0.19)	0.53 (0.28)	−0.08 (0.44)	−0.60 (0.35)
Positional Error (e_p) **(in pixels)** Avg. (s.d.)	*mindtct*	2.27 (0.59)	2.87 (0.72)	3.86 (0.72)	4.55 (1.05)
	COTS-A	2.07 (0.55)	2.54 (0.61)	3.43 (0.67)	4.17 (0.73)
	COTS-B	2.11 (0.63)	2.75 (0.74)	3.80 (0.69)	4.54 (0.72)
	COTS-C	2.24 (0.64)	2.85 (0.79)	3.84 (0.91)	4.82 (2.02)
Avg. Orientation Error (e_θ) **(in rad.)** Avg. (s.d.)	*mindtct*	0.06 (0.04)	0.09 (0.07)	0.19 (0.14)	0.36 (0.30)
	COTS-A	0.03 (0.02)	0.04 (0.03)	0.06 (0.05)	0.13 (0.12)
	COTS-B	0.04 (0.02)	0.05 (0.03)	0.07 (0.06)	0.13 (0.12)
	COTS-C	0.03 (0.02)	0.04 (0.03)	0.07 (0.07)	0.14 (0.25)

Table 4: Robustness evaluation of four minutiae extractors (*mindtct*, and COTS A, B, and C) against different levels of noise (Dataset-B).

Noise Levels	Minutiae Extractor	Level 1	Level 2	Level 3	Level 4
Paired Minutiae / Ground Truth (P_i / M_i) Avg. (s.d.)	*mindtct*	0.75 (0.12)	0.63 (0.11)	0.42 (0.09)	0.24 (0.08)
	COTS-A	0.92 (0.14)	0.88 (0.16)	0.81 (0.17)	0.70 (0.07)
	COTS-B	0.78 (0.15)	0.74 (0.12)	0.64 (0.10)	0.51 (0.08)
	COTS-C	0.89 (0.14)	0.80 (0.11)	0.52 (0.09)	0.24 (0.08)
Spurious Minutiae / Ground Truth (D_i / M_i) Avg. (s.d.)	*mindtct*	0.14 (0.06)	0.18 (0.09)	0.27 (0.13)	0.28 (0.12)
	COTS-A	0.04 (0.04)	0.05 (0.04)	0.10 (0.08)	0.24 (0.18)
	COTS-B	0.03 (0.03)	0.04 (0.04)	0.09 (0.07)	0.17 (0.10)
	COTS-C	0.05 (0.05)	0.08 (0.06)	0.11 (0.08)	0.08 (0.08)
Missing Minutiae / Ground Truth (I_i / M_i) Avg. (s.d.)	*mindtct*	0.25 (0.12)	0.37 (0.14)	0.58 (0.13)	0.76 (0.12)
	COTS-A	0.08 (0.06)	0.12 (0.07)	0.19 (0.08)	0.30 (0.12)
	COTS-B	0.22 (0.09)	0.26 (0.09)	0.36 (0.10)	0.49 (0.12)
	COTS-C	0.11 (0.09)	0.20 (0.13)	0.48 (0.22)	0.76 (0.19)

Table 5: Performance comparison of the four minutiae extractors (*mindtct*, and COTS A, B, and C) in terms of average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae for fingerprint images with different levels of noise (Dataset-B).

all minutiae extractors. Tab. 7 presents the performance comparison in terms of average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae for images with different levels of motion blur. With increase in the motion blur levels, a much higher percentage of missed minutiae is observed compared to paired and spurious minutiae.

4 Conclusions

Minutiae extraction is one of the most critical component of an automatic fingerprint identification systems. We have presented a controlled and repeatable evaluation of one open-source and three COTS minutiae extractors. Our experiments involve five public domain databases with manually marked minutiae to determine minutiae detection and localization accuracies. A large synthetically generated database with controlled levels of image degradations allowed us to quantify the affects of noise and motion blur, on minutiae extraction performance. The open-source minutiae extractor (*mindtct*) is observed to produce lowest positional errors in public domain databases. However, it also generates a higher percentage

Motion Blur Levels	Minutiae Extractor	Level 1	Level 2	Level 3
Goodness Index Avg. (s.d.)	*mindtct*	0.76 (0.12)	0.40 (0.16)	−0.68 (0.24)
	COTS-A	0.90 (0.13)	0.48 (0.16)	−0.50 (0.25)
	COTS-B	0.81 (0.15)	0.51 (0.15)	−0.56 (0.17)
	COTS-C	0.88 (0.10)	0.46 (0.13)	−0.70 (0.26)
Positional Error (e_p) **(in pixels)** Avg. (s.d.)	*mindtct*	3.05 (0.19)	3.69 (0.37)	4.14 (0.35)
	COTS-A	3.13 (0.20)	3.73 (0.38)	4.09 (0.31)
	COTS-B	3.08 (0.22)	3.84 (0.47)	4.10 (0.40)
	COTS-C	3.11 (0.19)	3.88 (0.31)	4.27 (0.58)
Avg. Orientation Error (e_θ) **(in rad.)** Avg. (s.d.)	*mindtct*	0.02 (0.01)	0.06 (0.02)	0.10 (0.02)
	COTS-A	0.01 (0.00)	0.06 (0.02)	0.09 (0.02)
	COTS-B	0.01 (0.01)	0.04 (0.01)	0.10 (0.03)
	COTS-C	0.01 (0.00)	0.06 (0.01)	0.08 (0.02)

Table 6: Robustness evaluation of four minutiae extractors (*mindtct*, and COTS A, B, and C) against different degrees of motion blur (Dataset-B).

Motion Blur Levels	Minutiae Extractor	Level 1	Level 2	Level 3
Paired Minutiae / Ground Truth (P_i / M_i) Avg. (s.d.)	*mindtct*	0.90 (0.09)	0.73 (0.14)	0.26 (0.18)
	COTS-A	0.96 (0.08)	0.76 (0.15)	0.34 (0.16)
	COTS-B	0.93 (0.09)	0.78 (0.14)	0.30 (0.16)
	COTS-C	0.95 (0.07)	0.75 (0.15)	0.25 (0.17)
Spurious Minutiae / Ground Truth (D_i / M_i) Avg. (s.d.)	*mindtct*	0.04 (0.03)	0.06 (0.04)	0.20 (0.13)
	COTS-A	0.02 (0.01)	0.04 (0.03)	0.18 (0.11)
	COTS-B	0.05 (0.03)	0.05 (0.04)	0.16 (0.13)
	COTS-C	0.02 (0.02)	0.04 (0.03)	0.20 (0.12)
Missing Minutiae / Ground Truth (I_i / M_i) Avg. (s.d.)	*mindtct*	0.10 (0.04)	0.27 (0.08)	0.74 (0.26)
	COTS-A	0.04 (0.02)	0.24 (0.06)	0.66 (0.19)
	COTS-B	0.07 (0.02)	0.22 (0.05)	0.70 (0.24)
	COTS-C	0.05 (0.02)	0.25 (0.06)	0.75 (0.20)

Table 7: Performance comparison of the four minutiae extractors (*mindtct*, and COTS A, B, and C) in terms of average percentages of paired (P_i/M_i), spurious (D_i/M_i), and missing (I_i/M_i) minutiae for fingerprint images with different levels of motion blur (Dataset-B).

of spurious minutiae compared to COTS minutiae extractors, deteriorating its overall performance. COTS-A exhibits significantly high robustness against different levels of image noise and motion blur.

References

[Al08] Alonso-Fernandez, F. et al.: Dealing with sensor interoperability in multi-biometrics: The UPM experience at the Biosecure Multimodal Evaluation 2007. In: SPIE Defense and Security Symposium. pp. 69440J–69440J, 2008.

[Ar16] Arora, S. S.; Cao, K.; Jain, A. K.; Paulter, N. G.: Design and Fabrication of 3D Fingerprint Targets. IEEE Trans. on Info. Forens. and Secur., 11(10):2284–2297, 2016.

[CDJ05] Chen, Y.; Dass, S. C.; Jain, A. K.: Fingerprint quality indices for predicting authentication performance. In: ICAVBPA. Springer, pp. 160–170, 2005.

[CFM08] Cappelli, Raffaele; Ferrara, Matteo; Maltoni, Davide: On the operational quality of fingerprint scanners. IEEE Transactions on Information Forensics and Security, 3(2):192–202, 2008.

[Do09] Dorizzi, B. et al.: Fingerprint and On-Line Signature Verification Competitions. In: ICB. pp. 725–732, 2009.

[Fl15] Flanagan, P.: Minutiae Interoperability Exchange III (MINEX III). NIST, 2015.

[Gh13] Ghiani, L. et al.: LivDet 2013 Fingerprint Liveness Detection Competition 2013. In: ICB. pp. 1–6, 2013.

[JFN10] Jain, A. K.; Feng, J.; Nandakumar, K.: Fingerprint matching. Computer, 43(2), 2010.

[Ka03] Kang, H. et al.: A study on performance evaluation of fingerprint sensors. In: ICAVBPA. Springer, pp. 574–583, 2003.

[KTU13] Kayaoglu, M.; Topcu, B.; Uludag, U.: Standard fingerprint databases: Manual minutiae labeling and matcher performance analyses. arXiv preprint arXiv:1305.1443, 2013.

[Li16] Liu, Xinwei; Pedersen, Marius; Charrier, Christophe; Bours, Patrick; Busch, Christoph: The Influence of Fingerprint Image Degradations on the Performance of Biometric System and Quality Assessment. In: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the. IEEE, pp. 1–6, 2016.

[Ma02] Maio, D. et al.: FVC2002: Second Fingerprint Verification Competition. In: ICPR. volume 3, pp. 811–814, 2002.

[Ma04] Maio, D. et al.: FVC2004: Third Fingerprint Verification Competition. In: ICBA, pp. 1–7. 2004.

[Ma09] Maltoni, D. et al.: Handbook of Fingerprint Recognition. Springer Science & Business Media, 2009.

[Na16a] National Institute of Standards and Technology, Development of NFIQ 2.0, https://www.nist.gov/services-resources/software/development-nfiq-20.

[Na16b] National Institute of Standards and Technology (NIST) Biometric Image Software (NBIS), https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis.

[NI] NIST SD27 Latent Database, https://www.nist.gov/srd/nistsd27.cfm.

[Ni05] Nill, N. B.: Test Procedures for Verifying IAFIS Image Quality Requirements for Fingerprint Scanners and Printers. MITRE Tech. Rep. MTR 05B0000016, 2005.

[Ni06] Nill, N. B.: Test Procedures for Verifying Image Quality Requirements for Personal Identity Verification (PIV) Single Finger Capture Devices. MITRE, MTR 060170, 2006.

[No14] Novetta Biosynthetic Software, https://www.novetta.com/wp-content/uploads/2014/11/NOV_Biosynthetics_Overview-2.pdf.

[RCJ95] Ratha, N. K.; Chen, S.; Jain, A. K.: Adaptive Flow Orientation-based Feature Extraction in Fingerprint Images. Pattern Recognition, 28(11):1657–1672, 1995.

[Tu11] Turroni, F. et al.: Improving fingerprint orientation extraction. IEEE Transactions on Information Forensics and Security, 6(3):1002–1013, 2011.

[Wa14] Watson, C. et al.: Fingerprint Vendor Technology Evaluation. NIST Interagency Report 8034, 2014.

[Wi04] Wilson, C. et al.: Fingerprint Vendor Technology Evaluation 2003: Summary of Results and Analysis Report. NIST Research Report, 7123, 2004.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 117

De-duplication using automated face recognition: a mathematical model and all babies are equally cute

Luuk Spreeuwers [1]

Abstract: De-duplication is defined as the technique to eliminate or link duplicate copies of repeating data. We consider a specific de-duplication application where a subject applies for a new passport and we want to check if he possesses a passport already under another name. To determine this, a facial photograph of the subject is compared to all photographs of the national database of passports. We investigate if state of the art facial recognition is up to this task and find that for a large database about 2 out of 3 duplicates can be found while few or no false duplicates are reported. This means that de-duplication using automated face recognition is feasible in practice. We also present a mathematical model to predict the performance of de-duplication and find that the probability that k false duplicates are returned can be described well by a Poisson distribution using a varying, subject specific false match rate. We present experimental results using a large database of actual passport photographs consisting of 224 000 images of about 100 000 subjects and find that the results are predicted well by our model.

Keywords: De-duplication, face recognition, large database, binomial distribution

1 Introduction

De-duplication is defined as the technique to eliminate or link duplicate copies of repeating data. In biometrics, there are several applications for de-duplication. One application is the cleaning of databases to make sure there is only one record per subject. A second application is to prevent that a new sample is entered in the database as a new entry, while a record of the subject already exists. In this paper, we address the 2nd category and more specifically, the application where a person applies for a new passport. The aim is to detect if this person already has a passport under another name. Currently, in the Netherlands, there exists a highly secured database of approximately 20 million subjects. The aim of this research was to investigate if it is feasible to, using modern state of the art automated facial recognition, determine if a subject has an entry in the database under another name. The main challenge in this context is the size of the database. In order to make the de-duplication feasible, if the photograph of an applicant is compared to the complete database, this should result in few to no false duplicates, caused by so-called look-a-likes, and should return true duplicates with a high probability. De-duplication becomes feasible if in 7-9 out of 10 applications, no false duplicates would be generated, while in 99 out of 100 applications the number of false duplicates would be less than 10. The latter means that an official has to manually inspect up to 10 returned images from the database

[1] Biometric Pattern Recognition Group, Chair of Services, Cyber Security and Safety (SCS), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, Netherlands, l.j.spreeuwers@utwente.nl

to decide if they are actual duplicates or are caused by look-a-likes. Further in order to be effective, the probability to detect actual duplicates should at least be above 50% (every second duplicate detected). These requirements were drafted in consultation with the Dutch passport issuance institution as realistic requirements.

There is not much literature available on de-duplication in face biometrics. In [DR13], an investigative study is presented on de-duplication errors. Two types of errors are introduced: False de-duplication (FDD) which is a match with a look-a-like and False non-duplication (FND) which corresponds to a missed duplicated. They provide results on a database with 1 009 identities. In [Ya11], de-duplication based on facial feature points is reported on a database of Chinese ID cards with 60 000 entries and 100/100 duplicates detected with 8 false hits. The main subject of the paper is, however, the presentation of a face recognition method based on 105 facial feature points, and the part on de-duplication performance is very brief. Scalability is not investigated at all. There are some reports on the related subject of large-scale 1:N comparison, see e.g. [GP04, GN14], but they do not explicitly address de-duplication.

One of the aims of our research is to investigate scalability to large databases of millions of entries. The following research questions were therefore formulated:

1. Is S.O.T.A. automated face recognition good enough to reliably detect duplicates in database with a size of 20 million entries?
2. What are the settings and further requirements for effective du-duplication?
3. Can the performance of de-duplication be predicted using a model?

In order to answer these questions, we developed a model for the de-duplication performance based on the binomial and Poisson distributions and set up an experiment using a database with approximately 100 000 subjects and 230 000 images and two commercial, state of the art automated facial recognition systems.

The remainder of this paper consists of the following sections: in section 2 a mathematical model is presented that describes the probability on errors and the probability to detect duplicates in large databases. In section 3, an experiment using a large database of 100 000 subjects is presented to verify the model. Finally, conclusions are presented in section 4.

2 A mathematical model for detection of duplicates

2.1 Errors in common biometric systems

In its basic form, a biometric system compares two biometric traces, e.g. facial images, and produces a similarity score s that is higher if the images are more similar. The aim of the biometric system is to determine if the two traces originate from the same the same subject. The similarity score is compared to a threshold T and if $s \geq T$, the traces are

Trace origins	result	type of match
Same subject	$s < T$	False Non Match (FNM)
Same subject	$s \geq T$	True Match (TM)
Distinct subjects	$s < T$	True Non Match (TNM)
Distinct subjects	$s \geq T$	False Match (FM)

Tab. 1: Types of matches of a biometric comparison

classified as coming from the same subject if not, they are regarded as traces from two different subjects. For a comparison 4 cases can be distinguished as shown in Table 1.

The performance of a biometric system is represented by an ROC graph, which shows the True Match Rate (TMR) as a function of the False Match Rate (FMR) for varying threshold. The ROC shows the trade-off between the TMR and the FMR: if the FMR decreases, then the TMR decreases and if the FMR increases, then the TMR also increases. If we choose T such that a certain FMR is realised, then from the ROC, we can read the TMR of the face comparison system. This is important for biometric systems that are used for verification applications, e.g. at border control where the one trace is the digital photograph stored in the passport and the other is a live recorded image. If the comparison results in a score higher than the given threshold, the probability that this is a True Match is estimated by the TMR and the probability on a False Match is estimated by the FMR, and both can be read from the ROC. The ROC is typically obtained using a large dataset of facial images.

An example of an ROC is given in Figure 1.

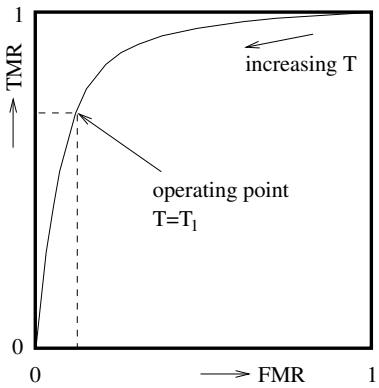


Fig. 1: ROC with operating point

A second common application of biometrics systems is the identification setting, where a single trace is compared to a list of traces of multiple subjects to check if the trace belongs to one of the subjects. We distinguish open set and closed set identification. In the former it is not known whether the owner of the trace is in the list of subjects, whereas in the latter case it is. Results are reported in the form of rank identification rates, where the rank-1 identification rate is an estimate of the probability that the subject in the list that results in

the highest score is the correct subject and rank-n that the correct subject is among the n highest scoring subjects in the list. In open set identification, also FNMR is reported and is also called False Negative Identification Rate (FNIR). Identification performance depends highly on the number of subjects in the list.

2.2 Performance of de-duplication

In [DR13], two types of de-duplication errors are distinguished: false de-duplication (FDD), i.e. the case that a duplicate is found while the corresponding trace in the database is actually not of the same subject as the probe trace, and false non-duplication (FND) where a trace of the same subject as the probe trace is present in the database, but not detected. These, however, apply to the case where one wants to build a database free of duplicates.

In our case, we want to detect duplicates of a facial photograph for a new passport application in a database. In order to make this feasible, we need to know the probability that a true duplicate (TD) is detected and the probability that the number of false duplicates (NFD) is below a certain threshold. For this we introduce the following measures:

Description	measure
Probability that a true duplicate is detected	$P(\text{TD})$
Probability on k false duplicates	$P(\text{NFD} = k)$
Probability that number of false duplicates is less than k	$P(\text{NFD} < k)$

Tab. 2: Measures for de-duplication, TD=True Duplicate, NFD=Number of False Duplicates

In the introduction we suggested that de-duplication is feasible in practice in the passport application if $P(\text{TD}) > 0.5$, $P(\text{NFD} = 0) > 0.7$ and $P(\text{NFD} < 10) > 0.99$.

2.3 A mathematical model for de-duplication

We assume that we have a facial image of a subject X and a large dataset of M images of which there are N_D duplicates and N images of other subjects. Furthermore, we assume that we have an automated face recognition (FR) system that compares two images, resulting in a score that is compared to a threshold T. The performance of the FR system is defined by its ROC, i.e. for a threshold T, we know the corresponding TMR and FMR.

If we compare the trace of X to all images in the database, then the probability that we detect a specific duplicate is given by the probability of a true match (α) when the trace is compared to a duplicate, i.e. it is estimated by the TMR obtained from the ROC.

$$P(\text{TD}) = \alpha \approx \text{TMR} \tag{1}$$

The probability on k false duplicates is modelled by a a series Bernoulli trials, where the probability on a false duplicate for a single comparison (β)is estimated by the FMR. The probability on k false duplicates is then given by the binomial distribution:

$$P(\text{NFD} = k) = \binom{n}{k} \beta^k (1-\beta)^{N-k} \tag{2}$$

This is the probability that k comparisons result in a score above T, while $N-k$ result in a score below T. The probability that less than k false duplicates are detected is then:

$$P(\text{NFD} < k) = \sum_{i=0}^{k-1} \binom{n}{k} \beta^k (1-\beta)^{N-k} \tag{3}$$

Note that an 1:N comparison is in practice not always described properly by N 1:1 comparisons, because FR systems may use various ways of score normalisation. For our derivations we ignore this effect.

Now, it can be shown that if N is very large and $N >> k$, then the binomial distribution can be approximated by the Poisson distribution [PP02]:

$$P(\text{NFD} = k) = \binom{n}{k} \beta^k (1-\beta)^{N-k} \approx \frac{1}{k!} \mu^k e^{-\mu} \tag{4}$$

Here, $\mu = N\beta$. Now this has an interesting implication if we want to predict the behaviour of de-duplication for varying database size N. If N increases by a factor λ, then if at the same time β (or the FMR) is decreased by a factor $\frac{1}{\lambda}$, the same probabilities result for $P(\text{NFD} = k)$ and $P(\text{NFD} < k)$!

The Poisson distribution has three different modes, depending on μ:

range of μ	behaviour as a function of k
$\mu \leq 1$	strictly decreasing
$1 < \mu \leq 5$	first going up, then down
$5 < \mu$	starting at nearly 0 going up then down

Tab. 3: Behaviour of the Poisson distribution as a function of μ

The three modes are also illustrated in Figure 2. Note that since k is an integer, the curves are not continuous.

Since we require $P(\text{NFD} = 0) > 0.7$, we need $\mu < 0.5$. As a matter of fact, we can calculate $P(\text{NFD} = 0)$ as a function of μ and likewise $P(\text{NFD} < k)$ as well. These relations are shown in Figure 3, where in the right figure $1 - P(\text{NFD} \leq 10)$ is plotted.

We can derive that for $P(\text{NFD} = 0) > 0.7$, we need $\mu < 0.36$, for $P(\text{NFD} = 0) > 0.9$, we need $\mu < 0.11$ and for all $\mu < 2$, $P(\text{NFD} < 10) >> 0.99$. Since $\mu = N\beta$, we can also

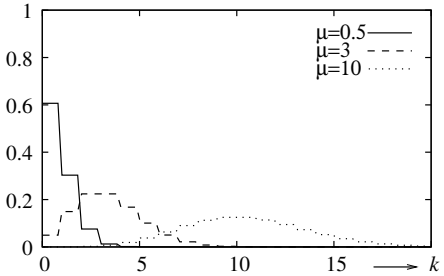


Fig. 2: Poisson distribution for various μ

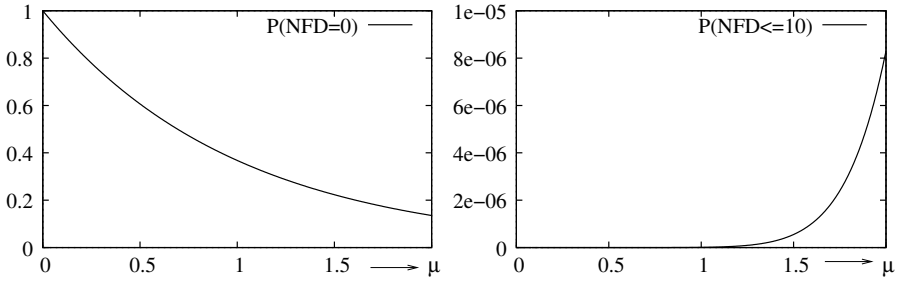


Fig. 3: $P(\text{NFD} = 0)$ and 1-$P(\text{NFD} \leq 10)$ as a function of μ

calculate the required β or FMR for a given dataset size. For various dataset sizes the required FMR values are given in Table 4.

N	β for $P(\text{NFD} = 0) = 0.9$	β for $P(\text{NFD} = 0) = 0.7$
1 000	$1.1 \cdot 10^{-4}$	$3.6 \cdot 10^{-4}$
100 000	$1.1 \cdot 10^{-6}$	$3.6 \cdot 10^{-6}$
200 000	$5.5 \cdot 10^{-7}$	$1.8 \cdot 10^{-6}$
10 000 000	$1.1 \cdot 10^{-8}$	$3.6 \cdot 10^{-8}$
20 000 000	$5.5 \cdot 10^{-9}$	$1.8 \cdot 10^{-8}$

Tab. 4: Required β or FMR for various dataset sizes

In conclusion, we can state that it is very well possible to predict the large scale behaviour of de-duplication using the Poisson distribution. There is, however, one catch: when we model the distribution $P(\text{NFD} = k)$ using the binomial distribution with constant β, we assume that for every subject, this β (or FMR) is the same. This, however, is not the case: some subjects are easier recognised than others and some subjects look more like each other than others. The used β is actually only the *average* β, $\bar{\beta}$ over all subjects. Thus β will vary per subject. In order to investigate the dependency of the results on the variation of β, we assumed that β would vary between $0.1\bar{\beta}$ and $1.9\bar{\beta}$ with a homogeneous distribution.

The probability on a certain number of false duplicated is thus calculated as:

$$P(\text{NFD} = k) = \int_{0.1\bar{\mu}}^{1.9\bar{\mu}} \frac{1}{k!}\mu^k e^{-\mu} d\mu \tag{5}$$

Where $\bar{\mu} = N\bar{\beta}$. Of course this is not the actual distribution of β, but it at least gives an indication of the effect of varying β for the different subjects. In Figure 4 the effect of varying μ (same as varying β, since $\mu = N\beta$) is shown.

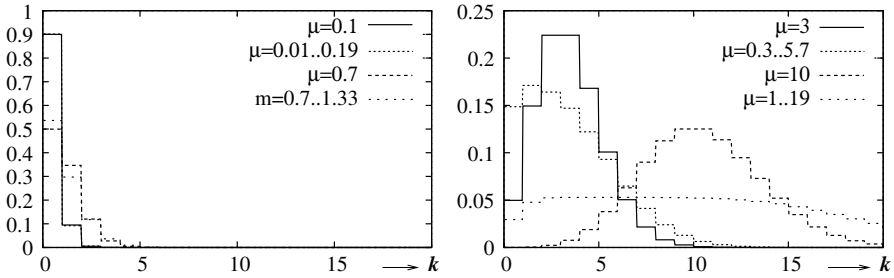


Fig. 4: Effect of non constant μ on $P(\text{NFD} = k)$, where $0.1\bar{\mu} < \mu < 1.9\bar{\mu}$

We can observe that for $\bar{\mu} = 0.1$, the effect is negligible (the curves for constant and varying μ coincide), for $\bar{\mu} = 0.7$ the peak at $k = 0$ is shifted up slightly and the tail becomes slightly longer. For larger $\bar{\mu}$, the peak of the curve $P(\text{NFD} = k)$ shifts to the left, while the whole curve becomes flatter and the right tail is longer.

Since we are interested in values of μ in the order of 0.1, we may expect that the subject specific variation in β has only small impact on the number of expected false duplicates.

3 An experiment on passport data

We set up an experiment with a database of passport photographs that was made available by the Ministry of Interior and Kingdom Relations of the Netherlands. Since strict privacy regulations apply to this database, the data could only be accessed in a highly secured environment and were only available for generating comparison scores and to a limited extend for visual inspection. In total the database consisted of 224 000 images of approximately 100 000 subjects. Of most subjects only two images were available, but of some more.

Using 2 commercial face recognition (FR) systems, all images of all subjects were compared to all other images, which would result in $50 \cdot 10^9$ scores. Due to time and space limitations, fewer scores were calculated. For the first system, 217 049 and for the second system 101 000 images were compared to all 224 000 images.

First the ROC for both FR systems were determined. They are not provided here, because their shape may reveal their origin. From Table 4, we can read the required FMRs (β)

that for databases of 200 000 and 20 000 000 images. For these settings the two facial recognition systems have a TMR as reported in Table 5.

Dataset size	$P(\text{NFD}=0)$	FMR	TMR system 1	TMR system 2
200 000	0.9	$5.5\cdot10^{-7}$	0.76	0.82
200 000	0.7	$1.8\cdot10^{-6}$	0.79	0.84
20 000 000	0.9	$5.5\cdot10^{-9}$	0.23	0.22
20 000 000	0.7	$1.8\cdot10^{-8}$	0.56	0.51

Tab. 5: FMR and TMR for two FR systems

From Table 5, we can see that for a dataset size of 200 000 the systems perform quite reasonably and allow for around 80% of the duplicates to be detected (4 out of 5). However, for a dataset of 20 000 000 the probability on detection a true duplicate drops to barely above 50% if $P(\text{NFD}=0)=0.7$. Note that with a FMR of $5.5\cdot10^{-9}$ we are at the limit of statistical certainty, because we have only about $20-40\cdot10^{9}$ false positive scores available. Also some subjects had a very high number of false duplicates, upto a few hundreds. Therefore, we visually inspected the images of the concerning subjects. To our surprise, they appeared to be all of babies and toddlers and young children, see Figure 5. As one of the results of this research we can therefore state that all babies look equally cute for the used FR systems. Indeed, poorer performance of FR for children has been reported before, see e.g. [GN14].

Fig. 5: All babies are equally cute (images obtained from the www)

We repeated the experiment with only subjects of ages above 14 years old, the results of which are represented in Table 6.

Dataset size	$P(\text{NFD}=0)$	FMR	TMR system 1	TMR system 2
200 000	0.9	$5.5\cdot10^{-7}$	0.89	0.92
200 000	0.7	$1.1\cdot10^{-6}$	0.92	0.94
20 000 000	0.9	$5.5\cdot10^{-9}$	0.28	0.27
20 000 000	0.7	$1.1\cdot10^{-8}$	0.65	0.65

Tab. 6: FMR and TMR for two FR systems for subjects with age 14+

We now see that for a database size of 20 000 000, 7 out of 10 subjects return no false duplicates and almost 2 out of 3 true duplicates are found according to our mathematical model, which, according to our set criteria is acceptable.

To investigate if the mathematical model is valid, we compared the predicted behaviour at various settings with the actual behaviour. From the complete set of 224 000 images, we drew 3 sets of 100 000, 10 000, and 1 000 images respectively and determined the probability on k false duplicates for a FMR such that $\mu = N\cdot\text{FMR} = 0.1$ (Figure 6 on

the left), and $\mu = N \cdot \text{FMR} = 1$ (Figure 6 right). We also predicted the behaviour with the models described in equations 4 and 5. These are shown as the solid curves in Figure 6.

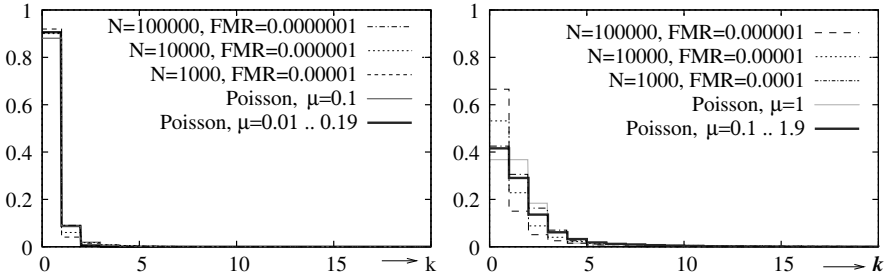


Fig. 6: Comparison of predictions by the mathematical model with actual measurements; for small μ, the model (drawn lines) match the measured results (various dashed/dotted lines) very well, while for larger μ the deviations are bigger

From the curves in Figure 6, we can observe that for small μ (left), the model predicts the behaviour very well and the behaviour for varying database sizes with fixed product $N \cdot \beta$ is replicated well. This means we can predict the behaviour for larger databases reliably. For larger μ, the accuracy of the prediction is less, but still the basic behaviour is characterised quite well (figure on the right). We can also observe that the model of Equation 5 for varying μ better predicts the behaviour than the Poisson distribution (Equation 4).

4 Conclusion

In this article we studied a specific de-duplication application where a subject applies for a new passport and we want to check if he possesses a passport already under another name. To determine this, a facial photograph of the subject is compared to all photographs of the national database of passports, in the Netherlands with a size of about 20 000 000. We investigate if state of the art facial recognition is up to this task and find that for a database of this size, duplicates can be detected with a probability of 65% (about 2 out of 3 duplicates is detected), while in 70% of all cases no false duplicates are reported and in more that 99% of all applications fewer than 10 false duplicates. This means that de-duplication using automated face recognition is feasible in practice.

We developed a mathematical model to predict the performance of de-duplication and find that the probability that k false duplicates are returned can be described well by a Poisson distribution using a varying, subject specific false match rate. An interesting and very useful property of the Poisson model is that if the database size increases N with a factor λ, the same behaviour is obtained provided the threshold for the FR system is chosen such that the FMR decreases with a factor $\frac{1}{\lambda}$, i.e. the product $N \cdot \text{FMR}$ remains constant.

Finally, we found that the used FR systems cannot distinguish small infants very well: for them all baby faces are equally cute.

References

[DR13] DeCann, B.; Ross, A.: De-duplication errors in a biometric system: An investigative study. In: 2013 IEEE International Workshop on Information Forensics and Security (WIFS). pp. 43–48, Nov 2013.

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

[GP04] Grother, P.; Phillips, P. J.: Models of large population recognition performance. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. volume 2, pp. II–68–II–75 Vol.2, June 2004.

[PP02] Papoulis, A.; Pillai, S.U.: Probability, random variables, and stochastic processes. McGraw-Hill electrical and electronic engineering series. McGraw-Hill, 2002.

[Ya11] Yang, Xiaoli; Su, Guangda; Chen, Jiansheng; Su, Nan; Ren, Xiaolong: Large Scale Identity Deduplication Using Face Recognition Based on Facial Feature Points. In (Sun, Zhenan; Lai, Jianhuang; Chen, Xilin; Tan, Tieniu, eds): Biometric Recognition: 6th Chinese Conference, CCBR 2011, Beijing, China, December 3-4, 2011. Proceedings. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 25–32, 2011.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 127

Multi-scale facial scanning via spatial LSTM for latent facial feature representation

Seong Tae Kim[1], Yeoreum Choi[1], Yong Man Ro[2]

Abstract: In the past few decades, automatic face recognition has been an important vision task. In this paper, we exploit the spatial relationships of facial local regions by using a novel deep network. In the proposed method, face is spatially scanned with spatial long short-term memory (LSTM) to encode the spatial correlation of facial regions. Moreover, with facial regions of various scales, the complementary information of the multi-scale facial features is encoded. Experimental results on public database showed that the proposed method outperformed the conventional methods by improving the face recognition accuracy under illumination variation.

Keywords: Face recognition, facial feature representation, spatial LSTM, deep learning

1 Introduction

In the past few decades, automatic face recognition has been an important vision task for many applications such as video surveillance and biometric identification [JRP04, CRP12, KKR16a]. For biometric identification, it is important to extract discriminative features which discriminate inter-person differences while being robust to intra-personal variations (e.g. illumination variations) [DCTD16].

As recent progress of deep learning, convolutional neural networks (CNN) have shown outstanding performance on many fields of computer vision such as image classification [KSH12, DCTD16], object detection [RHGS15], and action recognition [JXYY13]. Recently, the CNN has also been used to solve face recognition problems by learning latent and discriminative features [PVZ15, CKR16, KKR16b]. Generally, the CNN is comprised of one or more convolutional layers with a subsampling layer and followed by one or more fully-connected layers. In the convolutional layer, the filters slide over input images with convolutional operation to encode local image features. The neurons of feature maps obtained by convolution layer are connected to neurons of the fully-connected layer. In other words, spatial information extracted from local regions is simply aggregated to construct the image features. However, there are spatial relationships in facial local regions, which could not be encoded in the conventional CNN framework for face recognition [PVZ15, CKR16].

[1] School of Electrical Engineering, KAIST, Daejeon 34141, Republic of Korea, Email: {stkim4978, cyr0703}@kaist.ac.kr. Both authors are equally contributed to this manuscript.
[2] School of Electrical Engineering, KAIST, Daejeon 34141, Republic of Korea, Email: ymro@kaist.ac.kr. Corresponding author

In this paper, we propose a novel face recognition framework using deep network to solve the abovementioned limitation of the conventional CNN for face representation. To exploit the spatial relationships of facial local regions, we devise a long short-term memory (LSTM) network with which the whole face is scanned sequentially. The LSTM network originally introduced for sequence learning [HS97, GMH13, KKR16a]. It incorporates memory cells with three control gates (i.e., input, forget, output). The memory cells can store, modify, and access an internal state to learn long-term dependencies [BSF94]. In the proposed method, spatial long short-term memory has been devised to learn spatial dependencies of facial local features extracted from facial local regions. The contributions of this paper are summarized as followings: 1) A novel framework has been devised to encode latent facial features from spatial relationship of facial local regions. First, the facial local features are encoded by the CNN. Then the each facial local feature is used to construct facial latent spatial relationship-feature by scanning the whole face image. In other words, the face is scanned by the spatial LSTM network to learn relationship and dependencies of spatially sequential facial local regions. The memory cells of the spatial LSTM enable the proposed deep network to discover latent relationship of facial local regions. 2) The effectiveness of the proposed framework has been validated on the public face database. By the experiments, it is verified that the proposed method is robust to extract facial features under illumination variation. Moreover, the performance of face recognition could be further improved with multi-scale spatial long short-term memory, which combines latent facial features learned from multi-scale facial local regions.

The rest of this paper is organized as follows: The proposed latent facial feature representation using facial scanning is described in Section 2. Face recognition with multi-scale facial scanning is explained in Section 3. Section 4 presents and discusses experimental results. Finally, Section 5 provides concluding remarks.

2 Proposed latent facial feature representation by spatial LSTM

Figure 1 shows the overview of the proposed latent facial feature representation. The proposed method consists of facial local feature representation and spatial LSTM network. To learn the proposed latent facial feature representation, each face image is divided into regions horizontally and vertically, as shown Fig. 1. The objective of spatial LSTM is to learn relationship and dependencies of spatially sequential facial local regions. The spatial LSTM network consists of horizontal LSTM networks and vertical LSTM network. For horizontal scan, we divide face evenly into N_h parts with overlapping between two eye centers. The two eye centers are located based on facial landmark detection method [AZCP14]. For vertical scan, N_v horizontal patch sets are evenly divided between an eye corner and a lip corner. Eye corner and lip corner are also located by the facial landmark detection method. In this way, we can acquire $N_h \times N_v$ facial local patches from a face image (as shown in Fig. 1).

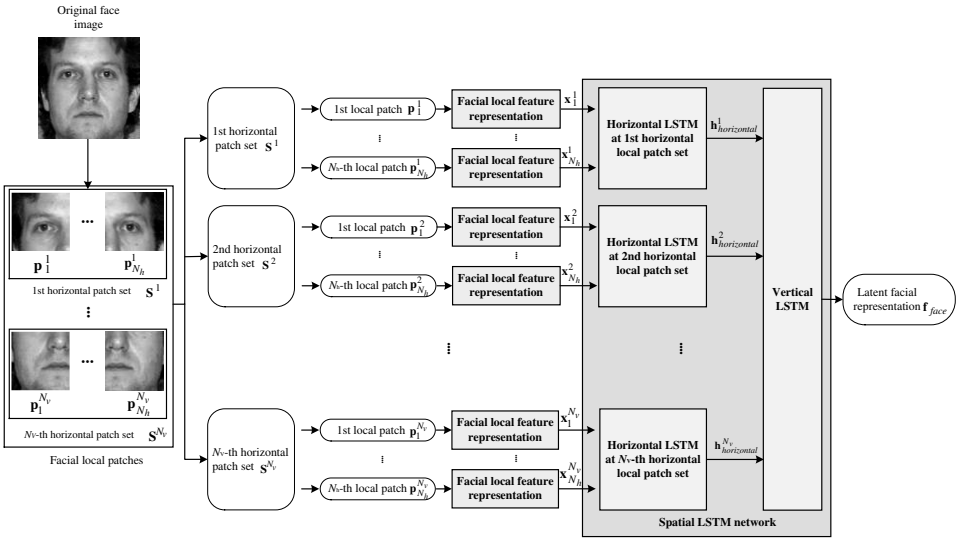


Fig. 1. Overall framework of the proposed latent facial feature representation. It consists of facial local feature representation and spatial LSTM.

The facial local features such as texture and shape are encoded by a CNN. The facial local features are used for input sequences of a spatial LSTM network. Let $\mathbf{F}^m = \left\{\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots \mathbf{x}_{N_h}^m\right\}$ denotes facial local features, which are extracted from the m-th horizontal patch set $\mathbf{S}^m = \left\{\mathbf{p}_1^m, \mathbf{p}_2^m, \ldots \mathbf{p}_{N_h}^m\right\}$ where $m = 1, 2, \ldots, N_v$. $\mathbf{p}_n^m$ denotes the n-th local patch in m-th horizontal patch set and $\mathbf{x}_n^m$ is the facial local feature encoded from $\mathbf{p}_n^m$.

We employ bidirectional LSTM to consider both directions in face scanning as:

$$\mathbf{h}_{fwd,n}^m = LSTM_{fwd}(\mathbf{x}_n^m, \mathbf{h}_{fwd,n-1}^m), \tag{1}$$

where $LSTM_{fwd}(\cdot)$ denotes a function which performs the operation of the LSTM layer in forward direction and $\mathbf{h}_{fwd,n}^m$ is the hidden state of the forward LSTM at n-th local patch in m-th horizontal patch set.

$$\mathbf{h}_{bwd,n}^m = LSTM_{bwd}(\mathbf{x}_n^m, \mathbf{h}_{bwd,n+1}^m), \tag{2}$$

where $LSTM_{bwd}(\cdot)$ denotes a function which performs the operation of the LSTM layer in backward direction and $\mathbf{h}^{m}_{bwd,n}$ is the hidden state of the backward LSTM at n-th local patch in m-th horizontal patch set. Then horizontal feature $\mathbf{h}^{m}_{horizontal}$ encoded at m-th horizontal patch set is represented as

$$\mathbf{h}^{m}_{horizontal} = [\mathbf{h}^{m}_{fwd,N_h}, \mathbf{h}^{m}_{bwd,1}]. \tag{3}$$

The vertical sequence acquired from horizontal LSTM network $\mathbf{h}_{horizontal} = \left\{ \mathbf{h}^{1}_{horizontal}, \mathbf{h}^{2}_{horizontal}, \ldots, \mathbf{h}^{N_v}_{horizontal} \right\}$ is used for the vertical LSTM to encode the facial feature vector $\mathbf{f}_{face}$ as followings:

$$\mathbf{f}_{fwd,m} = LSTM_{fwd}(\mathbf{h}^{m}_{horizontal}, \mathbf{f}_{fwd,m-1}), \tag{4}$$

$$\mathbf{f}_{bwd,m} = LSTM_{bwd}(\mathbf{h}^{m}_{horizontal}, \mathbf{f}_{bwd,m+1}), \tag{5}$$

$$\mathbf{f}_{face} = [\mathbf{f}_{fwd,N_v}, \mathbf{f}_{bwd,1}], \tag{6}$$

where $\mathbf{f}_{fwd,m}$ is the hidden state of the forward LSTM at m-th horizontal feature and $\mathbf{f}_{bwd,m}$ is the hidden state of the backward LSTM at m-th horizontal feature. Consequently, both horizontal LSTM networks and vertical LSTM network can learn gradual changes with respect to facial local distributions.

Fig. 2. Various scales of local patches at eye region for multi-scale facial scanning.

3 Face recognition with multi-scale facial scanning

From aforementioned spatial LSTM network which consists of horizontal LSTM networks and vertical LSTM network, we obtain a facial feature vector. By changing the

size of local region which is used to encode facial local features, the various facial feature vectors can be encoded in the spatial LSTM network. Therefore, combining these multi-scale facial features obtained from facial scanning, the complementary information could be encoded for face recognition. For this purpose, local patches are extracted with various sizes for considering multi-scale local features. In details, we acquire local patches with scale factor α which determines the ratio of the size of local region to the size of whole face image as shown in Fig. 2. Finally, the facial feature vectors extracted from various scales are combined as followings:

$$\mathbf{f}_{multiscale} = [\mathbf{f}_{face,1}, \cdots, \mathbf{f}_{face,\alpha}, \cdots, \mathbf{f}_{face,N_s}], \tag{7}$$

where $\mathbf{f}_{face,\alpha}$ denotes the facial feature vector obtained from facial scanning using local patch size of $\frac{1}{\alpha}$ and N_s denotes the number of multi-scale approach. Finally, a feature vector $\mathbf{f}_{multiscale}$ is used for face recognition. For the face recognition, 1-nearest neighborhood classifier is used based on Euclidean distance.

4 Experiments

4.1 Experimental conditions

To verify the proposed method, we performed experiments with the publicly available CMU Multi-PIE database which was collected from the face images under 20 illumination conditions (as seen in Fig. 3) [GMCKB10]. Particularly, the effectiveness of the proposed method under environment variation (i.e., varying illumination conditions) was investigated in this paper. We followed the experimental protocol in [CKR16] as followings. Among 337 subjects, we used mutually exclusive setting between the training set and the test set for evaluating the proposed method. The first 200 subjects were used for the training set and the remaining 137 subjects were used for the test set. In the case of test phase, the gallery images were set with only one frontal illumination condition and the probe images were chosen with other varying illumination conditions. In other words, the face images with 19 other illumination conditions of the database were included in the probe images. The number of gallery and probe images was 137 and 2,603, respectively.

In the experiment, N_h and N_v were set to 7 for cropping facial local regions. Each cropped facial local region was resized to 32×32 pixels. To extract feature vectors from facial local regions, the CNN structure [SKR15, CKR16] which consisted of three convolutional layers with a max-pooling layer, and two fully-connected layers was

adopted. For the case of LSTM network, each forward and backward LSTM layer has 512 memory cells, respectively. The proposed deep network was implemented by using the Keras framework with Theano backend [Ch15]. To avoid over-fitting, fully-connected layers and LSTM layers were constrained using drop out [SHKSS14].

Fig. 3. Example face images from CMU Multi-PIE under 20 different illumination conditions.

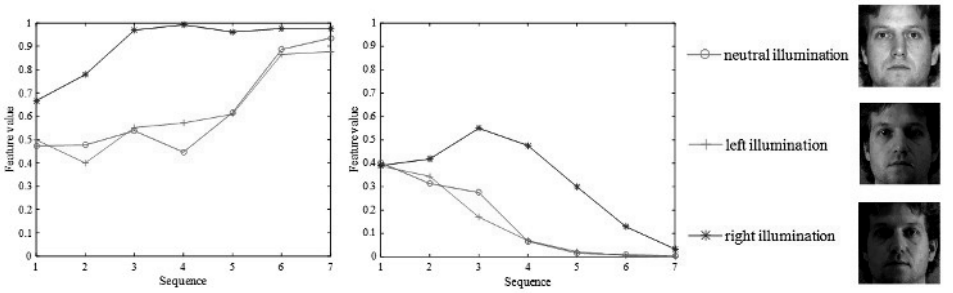


Fig. 4. Examples of feature changes according to sequential inputs of horizontal LSTM network. For visualization purpose, one of the feature value was selected from $\mathbf{h}^1_{horizontal}$ and normalized to [0, 1]. (Red (o): neutral illumination, green (+): left illumination, blue (*): right illumination)

4.2 Analysis of spatial LSTM for each illumination

Figure 4 shows the process of feature changes according to sequential inputs. The direction of input sequences was left to right. Each figure represents specific feature value of output feature vector from the LSTM network. One of the feature values obtained from the first horizontal LSTM network was used for the visualization. There were three face examples with different illumination conditions, which were neutral, left, and right illumination. As shown in each figure, the feature values of face images

showed similar changes under neutral and left illumination. On the contrary, the values of face images showed different tendency under right illumination. Nevertheless, all the values converged to feature values as bright parts of face images were put into LSTM network. These results indicated that the proposed method had the ability to store important information and forget noisy information, which resulted in encoding discriminative features under illumination variations.

Method	Accuracy
LBP [AHP06]	68.33%
GradientFace [ZTFS09]	84.75%
Weber-Face [WLYL11]	90.47%
VGG-Face [PVZ15]	85.06%
Two-step CNN [CKR16]	96.24%
Proposed method (α=2)	**96.73%**
Proposed method (multi-scale)	**98.08%**

Table 1. Accuracy of face recognition of the proposed method on CMU Multi-PIE database.

4.3 Face recognition performance under illumination variations

Table 1 shows the face recognition accuracy of the proposed method for CMU Multi-PIE database. For the comparison, local binary pattern (LBP) [AHP06], GradientFace [ZTFS09], Weber-face [WLYL11], VGG-face [PVZ15], and two-step CNN [CKR16] were used. The LBP was one of the popular approaches for local texture feature representation. The GradientFace and Weber-face were photometric normalization-based approaches for illumination variation. The VGG-face was CNN model learned from large scale celebrity face images. In this study, the pre-trained VGG-face model was fine tunned on the CMU MultiPIE database. The two-step CNN was the CNN-based approach which compensated illumination effects. As shown in the table, the proposed latent spatial facial feature representation achieved the accuracy of 96.73% at α=2. It outperformed other methods. This result indicated that encoding spatial sequential relationships between facial local regions was useful for face representation.

For the multi-scale facial scanning, N_S was set to 5. The proposed multi-scale approach achieved 98.08% accuracy by combining the multi-scale facial features obtained from various size of facial local regions. It was mainly attributed to the fact that the multi-scale approach could exploit the complementary information of multi-scale spatial long

short-term memory.

5 Conclusions

In this paper, we proposed the multi-scale facial scanning via spatial LSTM for latent facial feature representation. By scanning the face using the spatial LSTM network, the proposed method could exploit the relationship of the facial local regions. The experimental results with CMU Multi-PIE dataset showed that sequential relationships of facial local regions encoded by spatial LSTM network were useful in face recognition under illumination varioations. It was mainly attributed to the fact that important information was stored and noisy information was deleted by considering the spatial relationships of facial local regions in the spatial LSTM network. Moreover, by combining the complementary information obtained from multi-scale approaches, the accuracy of face recognition could be further improved in the proposed method.

6 Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIP) (No. 2015R1A2A2A01005724).

References

[AHP06] Ahonen, T.; Hadid, A.; Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12): 2037-2041, 2006.

[AZCP14] Asthana, A.; Zafeiriou, S.; Cheng, S.; Pantic, M.: Incremental face alignment in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1859-66, 2014

[BSF94] Bengio, Y.; Simard, P.; Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Networks, 5(2): 157-166, 1994.

[Ch15] Chollet, F: Keras: Theano-based deep learning library. 2015.

[CRP12] Choi, J. Y.; Ro, Y. M.; Plataniotis, K. N.: Color local texture features for color face recognition. IEEE Transactions on Image Processing, 21(3): 1366-1380, 2012.

[CKR16] Choi, Y.; Kim. H.-I.; Ro, Y. M.: Two-step Learning of Deep Convolutional Neural Network for Discriminative Face Recognition under Varying Illumination. In: Electronic Imaging. 2016.

[DCTD16] Ding, C.; Choi, J.; Tao, D.; Davis, L.: Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(3): 518-5531, 2016.

[GMH13] Graves, A.; Mohamed, A.-r.; Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6645-6649. 2013.

[GMCKB10] Gross, R.; Matthews, I.; Cohn, J.; Kanade, T.; Baker, S.: Multi-pie. Image and Vision Computing, 28(5): 807-813, 2010.

[HS97] Hochreiter, S.; Schmidhuber, J.: Long short-term memory. Neural Computation, 9(8): 1735-1780, 1997.

[JRP04] Jain, A. K.; Ross, A.; Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology, 14(1): 4-20, 2004.

[JXYY13] Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 221-231, 2013.

[KKR16a] 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: IEEE International Conference on Biometrics: Theory, Applications, and Systems. 2016.

[KKR16b] Kim, S. T.; Kim, D. H.; Ro, Y. M.: Spatio-temporal representation for face authentication by using multi-task learning with human attributes. In: IEEE International Conference on Image Processing. pp 2996-3000. 2016.

[KSH12] Krizhevsky, A.; Sutskever, I.; Hinton, G. E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. pp 1097-1105. 2012.

[MH08] Maaten, L.V.D.; Hinton, G: Visualizing data using t-SNE. In: ournal of Machine Learning Research. pp. 2579-2605. 2008.

[PVZ15] Parkhi, O. M.; Vedaldi, A.; Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. 2015.

[RHGS15] Ren, S.; He, K.; Girshick, R.; Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp 91-99. 2015.

[SHKSS14] Srivastava, N; Hinton, G; Krizhevsky, A; Sutskever, I; Salakhutdinov, R: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1): 1929-1958. 2014.

[SKR15] Seo, J.-J.; Kim, H.-I.; Ro, Y. M.: Pose-robust and discriminative feature representation by multi-task deep learning for multi-view face recognition. In: IEEE International Symposium on Multimedia. pp 166-171. 2015.

[SZ14] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. 2014.

[WLYL11] Wang, B.; Li, W.; Yang, W.; Liao, Q.: Illumination normalization based on weber's law with application to face recognition. IEEE Signal Processing Letters, 18(8): 462-465, 2011.

[ZTFS09] Zhang, T.; Tang, Y. Y.; Fang, B.; Shang, Z.; Liu, X.: Face recognition under varying illumination using gradientfaces. IEEE Transactions on Image Processing, 18(11): 2599-2606, 2009.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 137

Texture-Based Eyebrow Recognition

Mehmet Ozgur Turkoglu [1], Tugce Arican [2]

Abstract: Recent studies show that eyebrows can be used as a biometric or soft biometric for recognition. In some scenarios such as partially occluded or covered faces, they can be used for recognition. In this paper, we study eyebrow recognition using texture-based features. We apply features which have not been used before for eyebrow recognition such as 3-patch local binary pattern and WLD (Weber local descriptor) features. Also, we use more conventional features such as uniform LBP (Local binary pattern) and HOG (Histograms of oriented gradients). Methods are tested on both small- and large-sized datasets of images taken from FRGC database. Our experiments show that using some of these texture-based features together increases the performance significantly. We achieved more than 95% recognition accuracy for left and right eyebrows.

Keywords: eyebrow recognition, texture-based, local descriptor, HOG, LBP, 3-patch LBP, 4-patch LBP, WLD

1 Introduction

Face recognition systems have become more popular due to the improvements in hardware and software components. However, although the hardware and software are more robust in handling problems such as low resolution or blurring, they are not capable of identifying people accurately in some cases. For instance, when a person's face is partially occluded or covered with a balaclava, face recognition systems may easily fail. For that kind of challenging cases, one can think to use only some region of the face instead of using the whole face. Compared to other parts of the face (eye, eyelids or lip etc.) eyebrows might be more robust to different conditions such as different facial expressions. In recent years, the periocular region (eye, eye surroundings, and with or without eyebrow) is studied extensively as an alternative to iris biometrics and it is effective to identify a person. The eyebrows cover a small area on the face but it still might be very effective as a periocular region to identify a person.

There are some difficulties regarding the eyebrow biometric. With makeup or cosmetic operations (e.g. botox, plucking, tattoos), eyebrow shape or texture can be altered. If the contrast between eyebrow hair and skin is low, it is difficult to detect them and extract enough features. However, these situations are not frequent, so the eyebrow modality still might be useful biometric.

Eyebrows provide high-contrast lines which make them very noticeable, but also they might be detected even from long distances unless they are very thin. Besides, being easily

[1] University of Twente, Faculty of EEMCS, Enschede, The Netherlands, m.o.turkoglu@student.utwente.nl
[2] University of Twente, Faculty of EEMCS, Enschede, The Netherlands, t.arican@student.utwente.nl

detected, they also provide high-quality features to discriminate people. Eyebrows have shape-related characteristics such as thickness, length, and arch type. To discriminate a person using this kind of basic properties, they have to be determined very precisely, so accurate eyebrow segmentation is needed. Since it is a challenging problem itself (Dong et al.[DW11] use shape features; however, they segment eyebrows manually.), using texture information is easier and more convenient for eyebrow biometrics. In this paper, we focus on texture features of the eyebrow and propose a method for eyebrow recognition using multiple texture features. Even though 3-patch LBP and WLD features were successfully applied in face recognition and promising results were acquired, they have never been used in eyebrow biometrics before. We are the first to apply 3-patch LBP, 4-patch LBP, and WLD features in eyebrow recognition. Our proposed method achieved 'state-of-the-art' performance. Moreover, our experiments illustrated that eyebrows themselves can be as effective as a whole periocular region for identification.

This paper is structured as follows. In Section 2, we review related work. In Section 3, we describe the methods in detail. Then, we mention datasets and experiments in Section 4. We give the results and discuss our findings and compare with previous works in Section 5 and draw a conclusion in Section 6.

2 Related Work

The research in eyebrow biometrics is still limited even though some previous works such as [SJS03] and [RCA14] showed that eyebrows are one of the most important features for face recognition systems. Dong et al.[DW11] investigated the shape-based eyebrow features for biometric recognition and gender classification. Shape-based features were extracted from manually segmented eyebrow regions and grouped into three categories such as global shape features, local area features, and critical point features. They used three different classifiers: Minimum Distance Classifier (MD), Linear Discriminant Analysis Classifier (LDA), and Support Vector Machine Classifier (SVM). Best recognition rates obtained were 89%, 91% on MBGC dataset and 78%, 72% on FRGC dataset for the left and the right eyebrows respectively. In [LL07], Li et al. proposed a HMM-based eyebrow recognition system. They used Fourier coefficients as features and constructed recognition system using K-means classifier and HMMs. The proposed method tested on a small database which includes 54 eyebrow images from 27 subjects. The method achieved the highest accuracy of 92.6%. [Xu12] integrated Radon transform and SPP (sparsity preserving projections) and showed the feasibility and validity of their method by conducting an experiment on BJUT eyebrow database (The highest recognition rate of 87.2% was reported.). [YHZ13] showed that eyebrows may have the potential to be used in the real world security applications. They designed an eyebrow recognition system via fast template matching and Fourier spectrum distance. The proposed method achieved an accuracy of 98.6% on the BJUT eyebrow database. [JS11] focused on partial face. They divided the face into 6 regions (strips) and used eyebrow, eye, nose, and lip strips for recognition. Features were extracted by using LBP, WHT-LBP(Walsh-Hadamard Transform-LBP), DCT-LBP, and DFT-LBP. Unlike previous works, they used a large dataset and followed NIST's FRGC EXPERIMENT 4 protocol which involves matching 8K uncontrolled images to

16K controlled images from the 466 subjects. The average rank-1 accuracy from the full face was calculated as 47.9%. On the other hand, average accuracy from eyebrow was observed as 31.7%. Although eyebrow strip covers 1/6 of the full face, the average accuracy decreases by only 16.2%. They concluded that the eyebrow can be used as a stand-alone biometric.

3 Methodology

3.1 Preprocessing

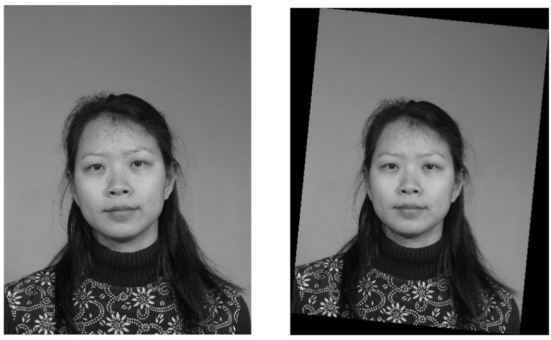
In our work, we use gray-scale images and use two different methods to obtain gray-level eyebrow images. In the first method, to align the image, an image is rotated in a way that left and right eye landmarks are aligned horizontally. Then, the image is cropped based on eyebrow landmark points (Leftmost, rightmost, topmost, and bottommost points are considered.). In the second method, we use fixed-eye coordinates. In this coordinate system, positions of eyes are constant. For instance, the xy-position of right and left eyes are (0,0) and (d,0) respectively (Here d is constant.). The image is transformed (rotation, scaling, and translation) in a way that left and right eye landmark points move to those specific positions in the space. Then, we place a fixed-size bounding box by incorporating eyebrow landmarks and crop the images. After applying both methods, the resulting images are resized to 36 by 90 pixels and used as an input image to extract features. In addition, we perform a histogram normalization in order to decrease the illumination effect before extracting features.

Fig. 1: Facial landmarks which are used in our work. There are 20 landmarks for an eyebrow and 1 landmark for an eye.

3.2 Texture Features

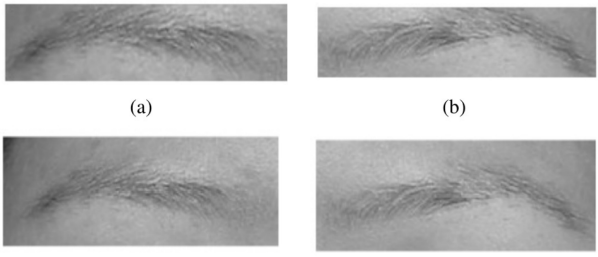
3.2.1 HOG (Histogram of Oriented Gradients) Features

The basic idea behind the HOG is local object appearance and shape can be described by local intensity gradient distributions or edge directions without knowing the edge positions [DT05]. In order to calculate the descriptor, the image is divided into small spatial regions and for each region, a local 1-D gradient histogram is calculated in a region. Then, histograms from each region are combined into a feature vector.



(a) (b)

Fig. 2: Aligning the face according to facial landmarks. (a) Original image. (b) Transformed image.



(a) (b)

(c) (d)

Fig. 3: Eyebrow images. (a)-(b) Images which are obtained by the first method. (c)-(d) Images which are obtained by the second method.

3.2.2 LBP (Local Binary Patterns) Features

The LBP is a descriptor of local spatial patterns and gray scale contrast. It was first introduced by Ojala et al.[OPH96]. Originally, each pixel in a 3-by-3 window is compared with the center pixel and labeled with 0 or 1 accordingly; then, the 8-bits binary code is obtained by concatenating all these labels in the window. The descriptor is created by calculating histograms of LBP codes in small regions. In this work, an uniform LBP which is an extension of the original LBP is used. In the uniform scheme, all the LBP codes which have more than 2 transitions (0 to 1 or 1 to 0) are assigned to one specific code.

3.2.3 Three-patch LBP Features

The three-patch LBP (TPLBP) descriptor is a different version of the LBP descriptor which was first introduced in [WHT08]. The way of producing each bit (0,1) in the code assigned to a single pixel differs. For each pixel in the image, a w by w patch centered on a pixel, and S additional patches distributed uniformly in a ring of radius r around the pixel are considered (see Figure 4). α is a parameter for the distance between two patches which are used at the same time. The value of a single bit is set according to which of the two

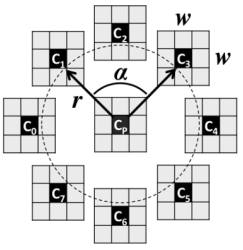
patches in the ring is more similar to the central patch. The resulting code has S bits per pixel. Simply the following formula is applied to each pixel.

$$TPLBP_{r,S,w,\alpha}(p) = \sum_{i}^{S} f(d(C_i, C_p) - d(C_{(i+\alpha) \bmod S}, C_p))2^i \quad (1)$$

Here $d(C_i, C_p)$ is a difference between pixel values (C_i, C_p). $f(x)$ is an unit step function.

$$f(x) = \begin{cases} 1 & x \geq 0 \\ 0 & otherwise \end{cases} \quad (2)$$

To calculate the 3-patch LBP descriptors, the same procedure with LBP features is applied. The image is divided into m by m small equal sized regions. For each region, a histogram is obtained by using 3-patch LBP codes. Then, these histograms are concatenated. The final histogram vector is called as 3-patch LBP descriptor.



$$\begin{aligned}
&\texttt{TPLBP}_{r,8,3,2}(p) = \\
&f(d(C_0, C_p) - d(C_2, C_p))2^0 + \\
&f(d(C_1, C_p) - d(C_3, C_p))2^1 + \\
&f(d(C_2, C_p) - d(C_4, C_p))2^2 + \\
&f(d(C_3, C_p) - d(C_5, C_p))2^3 + \\
&f(d(C_4, C_p) - d(C_6, C_p))2^4 + \\
&f(d(C_5, C_p) - d(C_7, C_p))2^5 + \\
&f(d(C_6, C_p) - d(C_0, C_p))2^6 + \\
&f(d(C_7, C_p) - d(C_1, C_p))2^7
\end{aligned}$$

Fig. 4: 3-patch LBP process with parameters $S = 8$, $\alpha = 2$ and $w = 3$. (Courtesy [WHT08])

3.2.4 Four-patch LBP Features

The four-patch LBP (FPLBP) was first introduced in [WHT08] as three-patch LBP. For each pixel in the image, two rings of radii r_1 and r_2 centered on a pixel, and S patches of size w by w spread out evenly on each ring are considered (see Figure 5). To produce a four-patch LBP code, two center symmetric patches in the inner ring with two center symmetric patches in the outer ring positioned α patches away along the circle (say, clockwise) are compared. One bit in each pixel's code is set according to which of the two pairs being compared is more similar. Thus, for S patches along each circle, there are $S/2$ center symmetric pairs which are the length of the binary codes produced. The formal definition of the four-patch LBP is following.

$$FPLBP_{r_1,r_2,S,w,\alpha}(p) = \sum_{i}^{S/2} f(d(C_{1,i}, C_{2,i+\alpha \ modS}) - d(C_{1,i+S/2}, C_{2,i+S/2+\alpha \ modS})2^i \quad (3)$$

After 4-patch LBP code is calculated for each pixel, the feature vector is created by applying the same procedure with LBP and 3-patch LBP.

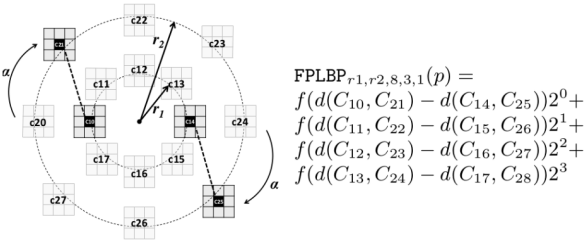


Fig. 5: 4-patch LBP process with parameters $S = 8$, $\alpha = 1$ and $w = 3$. (Courtesy [WHT08])

3.2.5 WLD Features

Weber local descriptor was first introduced in [Ch08]. It is inspired by Weber's Law[NE06] which is a perceptual law and simply states that the size of a just noticeable difference (ΔI) is a constant (k) proportion of the original stimulus value (I).

$$\frac{\Delta I}{I} = k \tag{4}$$

For instance, in a noisy environment, one must shout to be heard while a whisper works in a quiet room.

This descriptor consists of two components: differential excitation and orientation. Differential excitation for each pixel is computed as following.

$$\xi(x_c) = \arctan(\sum_{i=0}^{p-1} \frac{x_i - x_c}{x_c}) \tag{5}$$

here, x_c is the central pixel and x_i is the neighboring pixel. In order to prevent differential excitation of the pixel from increasing or decreasing too quickly, the arctan function is used as the excitation function (A sigmoid function can be used as well.). The orientation component is computed as following.

$$\theta(x_c) = \arctan(\frac{x_{right} - x_{left}}{x_{bottom} - x_{top}}) \tag{6}$$

here, x_{right} corresponds to right neighboring pixel of x_c. After differential excitation and orientation are computed, two suitable bins (one for excitation and one for orientation) are assigned to each pixel in the image, so each pixel is encoded with two numbers. Then, the image is divided into m by m small equal sized regions. For each region, 2-D histogram $WLD(\xi_j, \theta_t)$ is obtained and then, the 2D histogram $WLD(\xi_j, \theta_t)$ is further encoded into a 1D histogram H by reshaping. The final descriptor is created by concatenating the histogram (H) of each region (See Figure 6).

3.3 Score Functions

To compute the similarity between different feature vectors, L2 and χ^2 distances are used which show higher performance compared to L1 and normalized cosine distances.

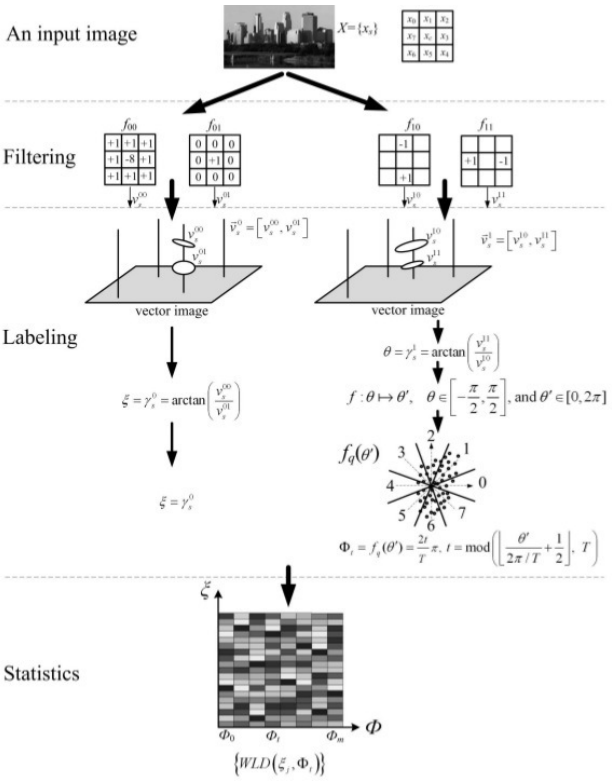


Fig. 6: Illustration of the computation of the WLD descriptor. (Courtesy [Ch10])

3.4 Score Fusion

Score-level fusion is used to fuse the information from different methods. Reduction of high-scores effect (RHE) normalization[He10], which is more robust compared to standard min-max normalization, is used. The normalized score x' is computed as following.

$$x' = \frac{x - X_{min}}{X^*_{mean} + X^*_{std} - X_{min}} \tag{7}$$

where x is the unnormalized score, X is the set of all the scores (genuine+impostor) and X^* is the set of genuine scores.

4 Experimental Setup

4.1 Dataset

In this work, the FRGC (Face Recognition Grand Challenge) v2.0[PF05] dataset was used. The images were taken in controlled and uncontrolled settings. Frontal images were taken

into two lighting conditions with two facial expressions(smiling and neutral) from different poses. In this work, only controlled images are used. In the first experiment, 500 images have been randomly chosen from 100 subjects (50 female, 50 male), 5 images each. In the second experiment, all the controlled images with facial landmark points are used, in total 12078 images from 568 subjects. The maximum number of images per subject is 70. In order to extract the eyebrow region, the locations of the facial landmark points have been obtained from DEST (Deformable Shape Tracking) facial landmarks. There are 199 landmarks in total and each eyebrow is encircled by 20 landmark points.

Fig. 7: Example (controlled) images from the FRGC dataset.

4.2 Experiments

We conducted 2 experiments. In the first experiment, we used 5 different feature extraction techniques (e.g HOG, LBP) and 2 different matching (preprocessing) techniques for left and right eyebrow images. We did not incorporate any learning (training) procedures. To compute the similarity between different feature vectors, we used Euclidean (L2) and χ^2 distances. Both verification and closed set identification tests were performed. For verification, a similarity matrix was created by taking the multiplicative inverse of each distance. In total, 1K genuine similarity scores and 123K impostor similarity scores were obtained. The performance was evaluated in terms of EER (Equal Error Rate). For identification, 3 images for each subject were randomly selected as the gallery image and the other 2 images were used as the probe image. The performance was evaluated in terms of rank-1 accuracy. In the second experiment, we used a large data-set. We followed the same procedure with the first experiment except we used only one matching technique and only one distance (L2 or χ^2) for each feature according to first experiment results. For identification, 200K genuine and 72 million impostor similarity scores were obtained. For identification test, half of the images were randomly selected as gallery image and the remaining images were used as probe images. In both experiments, information from different sources was fused in score level and the performances are reported in the same way.

5 Results & Discussion

The results of the experiments are listed in Table I and II and IV in terms of the equal error rate (EER) and in Table III and V in terms of the rank-1 accuracy. In order to obtain high performance, parameters of each method were optimized roughly in the first experiment and in the second experiment exactly the same parameters were used. The second matching

(preprocessing) technique gives better results (see Table I); therefore, for the rest of the experiments, only the second method was used. TPLBP and FPLBP features give lower EER with Euclidean (L2) distance; whereas, HOG and WLD features give lower EER with χ^2 distance; thus in the rest of the experiments, L2 distance is used for TPLBP and FPLBP features; χ^2 distance is used for the other features.

Our work shows that for eyebrow recognition using texture-based approach, transforming an image into a fixed-eye reference frame and using a fixed-size bounding box (method 2) is much better matching technique than using a variable-size bounding box according to eyebrow landmarks and re-sizing it to fixed size (method 1). Method 2 is more robust to inter-shape changes, so it provides impostor similarity distribution with lower variance. We think that the main reason that causes lower performance of method 1 is that eyebrow landmarks are not perfectly located on the boundary of the eyebrow (especially at the right and left tails of an eyebrow).

The uniform LBP descriptor outperforms the rest of the descriptors in terms of both EER and accuracy. Score fusion improves the performance. Even though FPLBP does not perform very well, it is effective when it is used with other features. According to results obtained, LBP and FPLBP together are quite robust and achieve the highest performance in most of the cases. The best result obtained in the first experiment is 6.3% EER, 97.5% rank-1 accuracy and 5.7% EER, 96.5% rank-1 accuracy for the left and the right eyebrows respectively. The results of the second experiment are compatible with of the first experiment. The best result obtained in the second experiment is 9.0% EER for both eyebrow and 96.2%, 95.3% rank-1 accuracy for the left and the right eyebrows respectively.

Results of the first experiment indicate that texture features are more robust compared to shape features which are studied in [DW11]. Dong et al.[DW11] tested their shape-based method on FRGC dataset with similar dataset size (400 samples, 100 subjects) and the best performance they achieved (7.0% EER) is worse than we obtained (5.7% EER). In addition, results show that eyebrows are robust to changes in illumination, pose and facial expression. Even though eyebrows cover significantly less area than the periocular region, only small degradation occurs in recognition performance. Mahalingam et. al.[MR13] studied the LBP-based periocular recognition and they conducted experiments on the FRGC dataset with similar dataset size and similar dataset split (50% gallery, 50% probe). They achieved 97.44% rank-1 accuracy which is only 1.2% higher than our highest rank-1 accuracy.

Distance	L2 Distance				χ^2 Distance			
	Left-1	Left-2	Right-1	Right-2	Left-1	Left-2	Right-1	Right-2
HOG	10.5	8.0	11.9	8.5	10.3	**7.4**	11.4	**8.0**
LBP	10.7	**7.2**	10.1	7.5	9.5	7.7	9.4	**7.0**
WLD	14.0	11.2	13.5	10.7	10.2	**9.6**	10.5	**8.3**
TPLBP	12.8	**9.8**	11.9	**8.4**	15.7	13.2	16.4	13.1
FPLBP	12.8	**8.6**	12.5	**8.8**	13.7	9.8	13.2	10.3

Tab. 1: Equal Error Rate (EER)'s of the first experiment.

Features	H+L	H+W	W+T	W+F	W+L	T+F	T+H	F+H	L+T	L+F
Left	**6.3**	8.0	8.1	7.4	8.0	7.7	7.1	6.7	7.1	6.6
Right	6.5	7.2	7.1	6.2	7.2	6.8	6.9	6.2	6.5	**5.7**

Tab. 2: Equal Error Rate (EER)'s of the first experiment using multiple features. For suitability, only first letters of the methods are shown (H:HOG, L:LBP, W:WLD, T:TPLBP, F:FPLBP).

Features	H	L	W	T	F	H+T	H+F	H+L	H+W	W+L	T+F	L+T	L+F
Left	92.5	**96.0**	92.5	92.5	88.5	95.0	94.0	**97.5**	96.5	96.5	94.5	96.0	96.0
Right	91.5	**95.0**	92.0	90.5	87.0	93.0	93.5	96.0	94.0	95.0	92.5	96.0	**96.5**

Tab. 3: Rank-1 accuracies of the first experiment.

Features	H	L	W	T	F	H+T	H+F	H+L	H+W	W+L	T+F	L+T	L+F
Left	11.1	**10.5**	12.9	12.6	11.2	10.4	9.6	9.9	10.6	10.6	10.3	10.2	**9.0**
Right	11.3	**10.1**	12.1	12.1	12.0	10.1	9.6	9.9	10.2	10.1	10.4	9.7	**9.0**

Tab. 4: Equal Error Rate (EER)'s of the second experiment.

6 Conclusion

In this work, we studied eyebrow recognition using several texture-based descriptors. These descriptors are HOG, uniform LBP, 3-patch LBP, 4-patch LBP, and WLD. We tested our methods on a large dataset which contains more than 12000 samples and the results we obtained show that LBP is more successful for eyebrow recognition problem. Also, we achieved better results by score fusion. The best result is obtained by using LBP and 4-patch LBP together, with 96.2% and 95.3% rank-1 accuracy for left and right eyebrows respectively. It suggests these texture-based features may be used for biometric recognition applications.

For the future work, we will test our method under non-ideal imaging conditions. In this work, we used pre-defined facial landmark points to extract the eyebrow regions, so in order to create a complete recognition system, we are planning to construct an automatic eyebrow detector.

References

[DW11] Dong, Yujie; Woodard, Damon L.: Eyebrow shape-based features for biometric recognition and gender classification: A feasibility study. Biometrics (IJCB), 2011 International Joint Conference on. IEEE, 2011.

[Ch10] Chen, Jie, et al.: WLD: A robust local image descriptor. IEEE transactions on pattern analysis and machine intelligence 32.9, pp: 1705-1720, 2010.

[WHT08] Wolf, Lior; Hassner, Tal; Taigman, Yaniv : Descriptor based methods in the wild. Workshop on faces in'real-life'images: Detection, alignment, and recognition. 2008.

[JS11] Juefei-Xu, Felix; Savvides, Marios : Can your eyebrows tell me who you are?. Processing and Communication Systems (ICSPCS), 2011 5th International Conference on. IEEE, 2011.

Features	H	L	W	T	F	H+T	H+F	H+L	H+W	W+L	T+F	L+T	L+F
Left	92.3	**95.0**	90.0	90.3	86.7	93.7	93.0	95.8	94.7	95.4	93.3	94.7	**96.2**
Right	90.8	**93.3**	90.8	89.0	84.4	92.0	91.6	94.5	93.4	94.4	92.2	93.8	**95.3**

Tab. 5: Rank-1 accuracies of the second experiment.

[PF05] Phillips, P.J.; Flynn, P.J.; Scruggs, T.; Bowyer, K.W.; Chang, J.; Hoffman, K.; Marques, J.; Min, J.; Worek, W.: Overview of face recognition grand challenge. IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[DT05] Dalal, Navneet; Triggs, Bill : Histograms of Oriented Gradients for Human Detection. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE, 2005.

[OPH96] Ojala, Timo; Pietikinen, Matti; Harwood, David: A comparative study of texture measures with classification based on featured distributions. Pattern recognition 29.1, pp: 51-59, 1996.

[YHZ13] Yujian, Li; Houjin, Li; Zhi, Cai : Human Eyebrow Recognition in the Matching-Recognizing Framework. Computer Vision and Image Understanding 117, pp: 170-181, 2013.

[He10] He, Mingxing, et al.: Performance evaluation of score level fusion in multimodal biometric systems. Pattern Recognition 43.5, pp: 1789-1800, 2010.

[NE06] Nutter, Forrest W.; Esker, Paul D. : The role of psychophysics in phytopathology: The Weber-Fechner law revisited. European Journal of Plant Pathology 114.2, pp: 199-213, 2006.

[MR13] Mahalingam, Gayathri; Ricanek, Karl : LBP-based periocular recognition on challenging face datasets. EURASIP Journal on Image and Video processing 2013.1, 2013:36.

[LL07] Li, Yujian; Li, Xingli : Hmm based eyebrow recognition. Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on. Vol. 1. IEEE, 2007.

[Xu12] Xu Xiaojun, Yang Xinwu, Li Yujian, and Yang Yuewei: Eyebrow recognition using radon transform and sparsity preserving projections. In Automatic Control and Artificial Intelligence (ACAI 2012), International Conference on, pages 1028â1033, March 2012

[SJS03] Sadr, Javid; Jarudi, Izzat; Sinha, Pawan : The role of eyebrows in face recognition. Perception 32.3, pp: 285-293, 2003.

[RCA14] Radji, Nadjet; Cherifi, Dalila; Azrar, Arab : Effect of eyes and eyebrows on face recognition system performance. Image Processing, Applications and Systems Conference (IPAS), 2014 First International. IEEE, 2014.

[Ch08] Chen, Jie, et al.: A robust descriptor based on weberâs law. Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 149

Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting

U. Scherhag[1], A. Nautsch[1], C. Rathgeb[1], M. Gomez-Barrero[1], R.N.J. Veldhuis[2], L. Spreeuwers[2], M. Schils[2], D. Maltoni[3], P. Grother[4], S. Marcel[5], R. Breithaupt[6], R. Raghavendra[7], C. Busch[1]

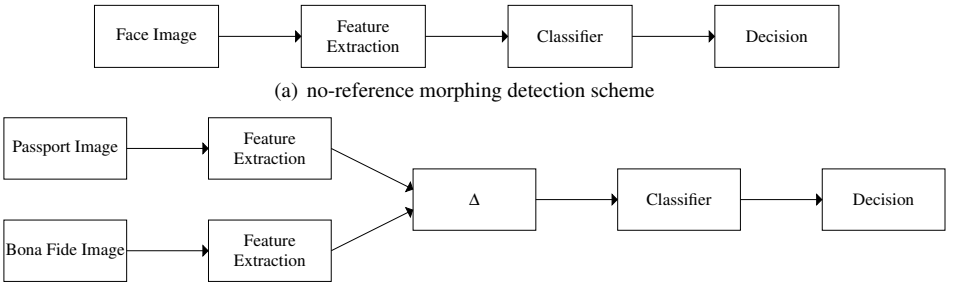
Abstract: With the widespread deployment of biometric recognition systems, the interest in attacking these systems is increasing. One of the easiest ways to circumvent a biometric recognition system are so-called presentation attacks, in which artefacts are presented to the sensor to either impersonate another subject or avoid being recognised. In the recent past, the vulnerabilities of biometric systems to so-called morphing attacks have been unveiled. In such attacks, biometric samples of multiple subjects are merged in the signal or feature domain, in order to allow a successful verification of all contributing subjects against the morphed identity. Being a recent area of research, there is to date no standardised manner to evaluate the vulnerability of biometric systems to these attacks. Hence, it is not yet possible to establish a common benchmark between different morph detection algorithms. In this paper, we tackle this issue proposing new metrics for vulnerability reporting, which build upon our joint experience in researching this challenging attack scenario. In addition, recommendations on the assessment of morphing techniques and morphing detection metrics are given.

Keywords: Biometrics, Morphing, Performance Reporting, Attack Detection

1 Introduction

Biometrics refers to the automated recognition of individuals based on their biological and behavioural characteristics [In12]. Due to the strong link between subjects and their biometric samples, the wide acceptance, and their user convenience, biometric systems become increasingly popular. Even though the security of biometric systems is increasing, recent research revealed a security gap to subvert the unique link between a biometric sample and its subject. By enrolling an artificial sample, generated by merging samples of two or multiple subjects in image or feature domain, the contributing subjects might be verified successfully against the manipulated reference. This can be done, for instance, in the passport application process, where in most countries the applicant brings his own printed photograph. This way, the unique link between individuals and their biometric reference data is annulled. The feasibility of such *morphing attacks* was first shown for face recognition systems [FFM14, FFM16] and most recently for fingerprint [FCM17]

[1] da/sec Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {ulrich.scherhag, andreas.nautsch, christian.rathgeb, marta.gomez-barrero, christoph.busch}@h-da.de
[2] Services, Cybersecurity and Safety Research Group, University of Twente, The Netherlands, {r.n.j.veldhuis, l.j.spreeuwers}@utwente.nl, m.schils@student.utwente.nl
[3] DISI, University of Bologna, Italy, davide.maltoni@unibo.it
[4] National Institute of Standards and Technology, patrick.grother@nist.gov
[5] Idiap Research Institute, Martigny, Switzerland, marcel@idiap.ch
[6] Bundesamt für Sicherheit in der Informationstechnik (BSI), Bonn, Germany, ralph.breithaupt@bsi.bund.de
[7] Norwegian Biometrics Laboratory, NTNU, Gjøvik, Norway raghavendra.ramachandra@ntnu.no



(a) no-reference morphing detection scheme

(b) differential morphing detection scheme

Fig. 1: Difference between no-reference and differential morphing detection schemes

and iris [RB17] recognition systems. The remainder of the paper will focus on the face case study, as it most widely studied and allows for a comprehensible visual explanation of the morphing process and the occurring difficulties.

To prevent the aforementioned morphing attacks, an automatic detection of morphs is required. Focusing on the workflow of a generic biometric system, two morph detection tasks can be distinguished: (1) detection during enrolment, e.g. the passport application process, where the detector processes a single image, referred to as no-reference morphing detection and depicted in Figure 1(a); and (2) detection at the time of authentication, e.g. the usage of Automated Border Control (ABC) gates at borders, where a live capture from an authentication attempt serves as additional source of information for the morph detector, referred to as differential morphing detection and depicted in Figure 1(b). Moreover, two attack scenarios can be distinguished: (i) an attacker could try to attack a fully-automated biometric system or (ii) a semi-automated system with human examiners in the loop. In the latter case, the role of subjects contributing to a morphed image might be asymmetric, i.e., some subjects might have to pass the human inspection while others have to pass biometric recognition systems.

The metrics and terminology defined in ISO/IEC 30107-3 on Presentation Attack Detection evaluation [In17] strongly relate to morphing attacks. However, those metrics only apply to one subject per attack. On the contrary, for morphing attacks success is achieved if multiple subjects bypass the system for a single sample, i.e., more than one biometric decisions have to be considered. Thus, only parts of this standard can be employed for evaluating morphing detection, while other metrics, e.g. the Impostor Attack Presentation Match Rate (IAPMR), need to be adapted.

Even if there exist some works dealing with morphing [FFM14, FFM16, Go17, RB17, FCM17] or morphing detection [RRB16, Sc17], no common understanding for morphing attacks and morphing detection has been developed yet. During the creation of morphing databases and the design of morphing detection algorithms we observed multiple pitfalls, which are summarized in this paper. In order to allow a common evaluation of the attack success rate, we propose new metrics, in particular the Mated Morph Presentation Match Rate (MMPMR) and Relative Morph Match Rate (RMMR). With this proposal,

we intend to spark a discussion within the research community and awaken the interest of the ISO/IEC biometrics standardisation committee to compose a comprehensive list of requirements that need to be taken into account when evaluating morphing attacks.

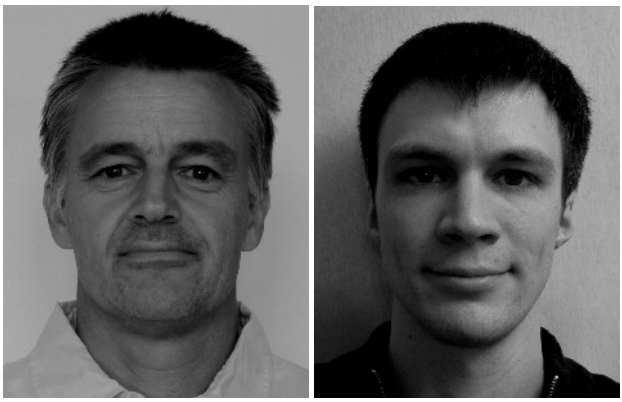
The remainder of this paper is organized as follows: In Section 2 we present observations from our work on creating morphed images, based on which diverse recommendations are given. Section 3 proposes metrics to evaluate the vulnerability of biometric systems to morphing attacks. In Section 4 recommendations for morphing detection and morphing detection evaluation are given.

2 Recommendations on Morph Generation

Morphing attack samples generated for research databases may differ from real world attack samples. In order to achieve significant evaluation results, a large number of attack samples has to be created, which can be achieved by automated methods. For the sake of realistic attack scenarios, four major factors have to be considered: (1) the morph quality, (2) the similarity of the morphed subjects, (3) the consistent quality of the database and (4) weights used to generate the morph. All of these factors will be discussed in detail in the subsequent paragraphs.

(1) The real world attacker has the option to spend much time on one single morph, which might reveal a higher quality compared to automatically generated images, depending on the goodness of the automatic morphing process and the skills of the attacker. Figure 2(c) shows a high quality morphed face image attack sample generated using FantaMorph [Ab17], whereas an example for a low quality morph is depicted in Figure 2(e). Both images are successfully verified against the contributing subjects, but Figure 2(e) contains a huge amount of obvious morphing artefacts which can be easily detected by human observers or common pattern recognition algorithms. As the attacker is willing to do his best to circumvent the system, the best conditions should be expected for the attacker. In particular, for face image morphing attacks in border control scenarios, the image needs to fulfill specific quality requirements, defined in [In05], in order to be accepted as a biometric passport sample. Thus, assuming the preserved chain-of-trust of the passport creation process, the appearance of obvious artefacts should be minimized during the morphing process. However, for evaluation purposes morphs with lower quality might be of interest as well. In order to achieve a significant evaluation on a database containing multiple quality levels of morphing attacks, a clear labeling of the data is mandatory. For automatically generated morphs, a consistent quality per algorithm is assumed. Manually generated morphs, however, might vary in quality, requiring a quality metric to ensure a proper labeling. The definition of the quality metric depends on the specific scenario and is not in the scope of this paper.

(2) As motivated in [Go17], morphs of two subjects yielding a high chance of both being positively matched, referred to as *lookalike morphs*, are more relevant than morphs of two subjects highly differing in appearance, referred to as *non-lookalike morphs*. One option is selecting subjects to be morphed according to the similarity score returned by a face recognition algorithm. However, in a real world scenario, realistic lookalike morphs are necessary to fool human experts [FFM16], e.g. when applying for an ID document. A high



(a) Subject 1 (b) Subject 2

(c) manual high quality morph (d) automated morph (e) low quality morph

Fig. 2: Differences between morphing qualities

number of non-lookalike or bad quality morphs might reduce the impact of the attack on the recognition system and at the same time artificially increase the detection performance of the morph detector. In order to achieve realistic combinations of subjects, a clear pre-categorization of the subjects according to soft-biometric attributes is recommended, e.g skin, hair, gender, or age in case of face images. The definition of the different categories falls out of scope of this work and needs to be addressed in further research. Employing such a classification scheme, the total number of morphs can be divided into subsets representing different relevance-classes.

(3) For verification purposes, training on images with different quality and resolution leads to a higher recognition accuracy and robustness to different scenarios. However, for morphing detection, it is important to obtain an equal quality for bona fide and morphed samples, in order to avoid bias towards different quality levels on the morphing detection algorithm. To illustrate this fact, Figure 3 depicts the impact of JPEG-compression on morphed face images. For quality estimation, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [MMB11] was employed. While the image of the originating subject,

(a) uncompressed face image of original subject
BRISQUE = 21.0

(b) uncompressed morphed face image
BRISQUE = 29.1

(c) compressed morphed face image
BRISQUE = 50.0

Fig. 3: Impact of JPEG-compression

(a) eye of manual morph

(b) eye of autom. morph

Fig. 4: Differences between different morphing techniques

as well as the morphed face image depicted in Figure 3(b) are uncompressed, Figure 3(c) is JPEG-compressed. Even if no visual difference can be observed between these images, the BRISQUE-score indicates a severe quality loss of the image due to the compression.

So far, we have observed that common machine learning algorithms, e.g. Support Vector Machines (SVM), trained on local image descriptors, e.g. Binarized Statistical Image Features (BSIF) [KR12] or Local Binary Patterns (LBP) [OPM02], might classify the images according to such compression artefacts or image quality differences, if present, instead of attributes related to the morphing process. The same applies for other artefacts introduced by some post-processing, e.g. rescaling, image-optimization or rotation. However, distinct artefacts might disappear if an image is printed and scanned.

The aforementioned quality issues are more likely to appear in automatically generated morphs, which are thus expected to differ in quality from manual morphs. Figure 2(c) depicts a manual morphed face image and Figure 2(d) an automatically generated morph. As emphasized in Figure 4(a) and 4(b), the automatically generated morph reveals small shadow artefacts which can be avoided in the manual morph.

(4) Another factor to consider is the weights of multiple subjects in a morph. This is a key factor in scenarios where humans are required to check the morph against a live subject. For border control scenarios, it could be feasible to generate a morphed image with a high weight for the accomplice applying for the passport. This way, the accomplice has a high chance of deceiving the officer at enrolment time and the criminal will still be able to be successfully verified by the ABC gate.

3 Assessment of Vulnerability to Morphing Attacks

To assess the vulnerability of a specific biometric system to morphing attacks, a standardized methodology is needed. In general, it is crucial to follow the guidelines proposed in [MW02], where it is recommended that all comparisons in one evaluation should be uncorrelated. In particular, the samples compared to the morphed face images should not be the same as the ones used for the morphing process, since such a comparison would ignore the natural biometric variance.

Regarding evaluation metrics, the Impostor Attack Presentation Match Rate (IAPMR) introduced in ISO/IEC 30107-3 on Presentation Attack Detection evaluation [In17] represents a standardized metric for attack success evaluation:

> **IAPMR:** in a full-system evaluation of a verification system, the proportion of impostor attack presentations using the same Presentation Attack Instrument (PAI) species in which the target reference is matched.

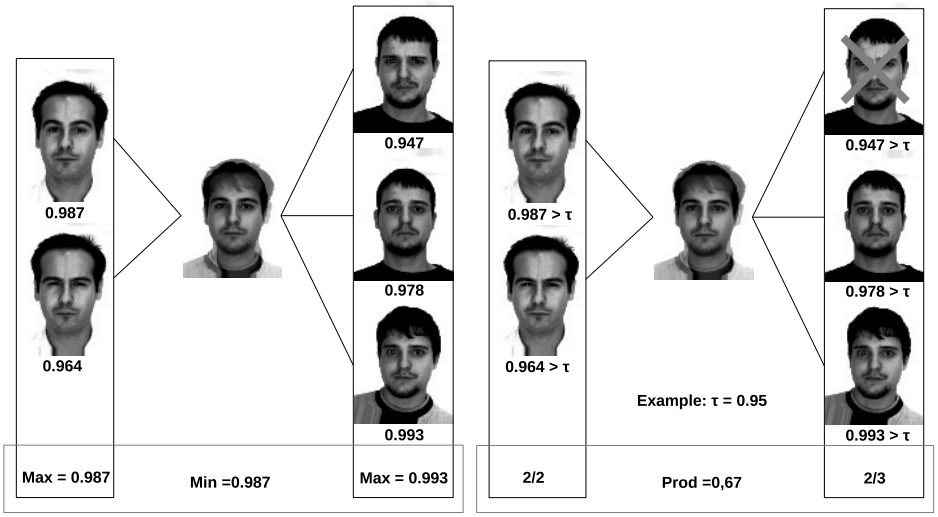
However, for the evaluation of morphing attacks, the aforementioned IAPMR metric presents some drawbacks, as a morphing attack can only be considered successful if all contributing subjects are successfully matched against the morphed sample. To avoid a confusion of the wording impostor, which is used for zero effort impostors, the comparison of a morphed sample to another independent sample of one contributing subject will be referred to as *mated morph comparison*. Motivated by the international standard ISO/IEC 30107-3 [In17], we propose a new metric for the evaluation of the impact of a morphing attack in a full-system evaluation, referred to as Mated Morph Presentation Match Rate (*MMPMR*).

If the recommendations of [MW02] are considered, only one mated morph comparison per subject is possible. As the morphing attack succeeds if all contributing subjects are verified successfully, only the minimum (for similarity scores) or maximum (for dissimilarity scores) of all mated morph comparisons of one morphed sample are of interest. The *MMPMR* for similarity scores is accordingly defined as:

$$MMPMR(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left\{ \left[\min_{n=1,\ldots,N_m} S_m^n \right] > \tau \right\}, \tag{1}$$

where τ is the verification threshold, S_m^n is the mated morph comparison score of the n-th subject of morph m, M is the total number of morphed images and N_m the total number of subjects contributing to morph m. The following examples are for similarity scores.

If, due to a lack of data, the recommendation in [MW02] is not met, and multiple samples of one subject are compared to one morphed image, there are two possibilities to adapt



(a) MinMax-MMPMR (Eq. (2)) (b) ProdAvg-MMPMR (Eq. (3))

Fig. 5: Examples for the computation of *MMPMR*

the metric. For smaller number of samples, multiple comparisons can be understood as multiple authentication attempts per subject. For instance, for face image morphing attacks in a border control scenario, the attacker is able to conduct several authentication attempts and will be successfully verified, as long as one attempt is above the threshold of the biometric system. Thus, the metric can be extended by only considering the maximum (for similarity scores) or minimum (for dissimilarity scores) over all mated morph comparisons of one subject, referred to as *MinMax-MMPMR* and depicted in Figure 5(a).

$$\textit{MinMax-MMPMR}(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left\{ \left(\min_{n=1,\ldots,N_m} \left[\max_{i=1,\ldots,I_m^n} S_m^{n,i} \right] \right) > \tau \right\}, \tag{2}$$

where I_m^n is the number of samples of subject n within morph m. *MinMax-MMPMR* also models the case where N_m subjects launch single attacks to several biometric authentication systems ($I_m^n = 1$).

However, for larger number of probe sample per subject, the MinMax approach is prone to falsely increase the number of accepted subjects. Thus, we propose a probabilistic interpretation, by calculating the proportion of accepted attempts per subject and multiply the probabilities of all contributing subjects (i.e., joint probability). The mated morph acceptance rate is calculated over all contributing subjects, referred to as *ProdAvg-MMPMR*, in analogy to the *MinMax-MMPMR* and depicted in Figure 5(b):

$$\textit{ProdAvg-MMPMR}(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left[\prod_{n=1}^{N_m} \left(\frac{1}{I_m^n} \cdot \sum_{i=1}^{I_m^n} \left\{ S_m^{n,i} > \tau \right\} \right) \right]. \tag{3}$$

MMPMR, as well as IAPMR, are directly dependent on the threshold τ of the biometric system. In order to achieve a more generalized metric, we propose to compute the dif-

ference between *MMPMR* or IAPMR and $1-FNMR$, respectively. The Relative Morph Match Rate (*RMMR*) is defined as follows:

$$RMMR(\tau) = 1 + (MMPMR(\tau) - (1 - FNMR(\tau))) = 1 + (MMPMR(\tau) - TMR(\tau)). \quad (4)$$

For presentation attacks as described in [In17], *MMPMR* can be replaced by *IAPMR*.

Figure 6 depicts different *RMMR* examples for combinations of distributions and thresholds. If *MMPMR* and $1-FNMR$ are equal sized, the *RMMR* will be 1 (Figure 6(a) and 6(e)). For a more restrictive threshold, the *RMMR* will decrease (Figure 6(b) and 6(c)), until the threshold reaches a point where the $FNMR$ increases (see Figure 6(d)). For a scenario in which all mated morph comparisons are rejected (as depicted in Figure 6(d)), the distribution of the comparison scores is not of interest. Even for a mated morph distribution far below the impostor comparisons, the *RMMR* would remain the same. If the score distribution of mated morphs is bigger than $1-FNMR$, the *RMMR* will be above 1 (see Figure 6(f)). Note that the latter "extreme" case is considered as unrealistic, since the *RMMR* is assumed to be upper-bounded by $1-FNMR$.

For security assessment scenarios, a morphed sample is a threat as soon as more than one subject is successfully verified against it. For these assessments only the two most successful subjects are considered for the *MMPMR* estimation. If more subjects are successfully verified against the morphed reference, the attack can be considered as more severe, thus the amount of successful mated morph comparisons should be reported as the weight of the attack.

4 Morphing Detection Performance Reporting

Multiple procedures for creating morphed images and/ or multiple morph detectors can be independently benchmarked employing the metrics defined in [In17]. In particular, the Attack Presentation Classification Error Rate (APCER) and the Bona Fide Presentation Classification Error Rate (BPCER) should be computed, and visualised in a Detection Error Trade-Off (DET) curve.

In addition, in order to achieve reproducible and comparable performance evaluations of morphing detection systems, for each procedure or detector a common comprehension of the training and testing methodology is needed. In general, the standards defined in ISO/IEC 19795-1 on biometric performance testing and reporting [In06] should be followed, e.g. a disjoint subdivision of the data into training and testing set. More specifically, a strict separation of the morphed samples with respect to the originating subjects is important, in order to avoid an unrealistic high detection performance. It should be noted, that one morphed sample is related to at least two subjects and each subject might contribute to several morphing samples.

As described in Section 2, when aiming to develop and test a robust detection algorithm, it is crucial to ensure, that the feature extractor is not based on artefacts present on low quality morphs. Otherwise, it is likely that a trained classifier might strongly rely on these specific artefacts. As a consequence, if different quality levels of morphed samples should be examined, these should be evaluated separately according the quality labels defined during the database creation process. For Example, if a morphing detection system is trained

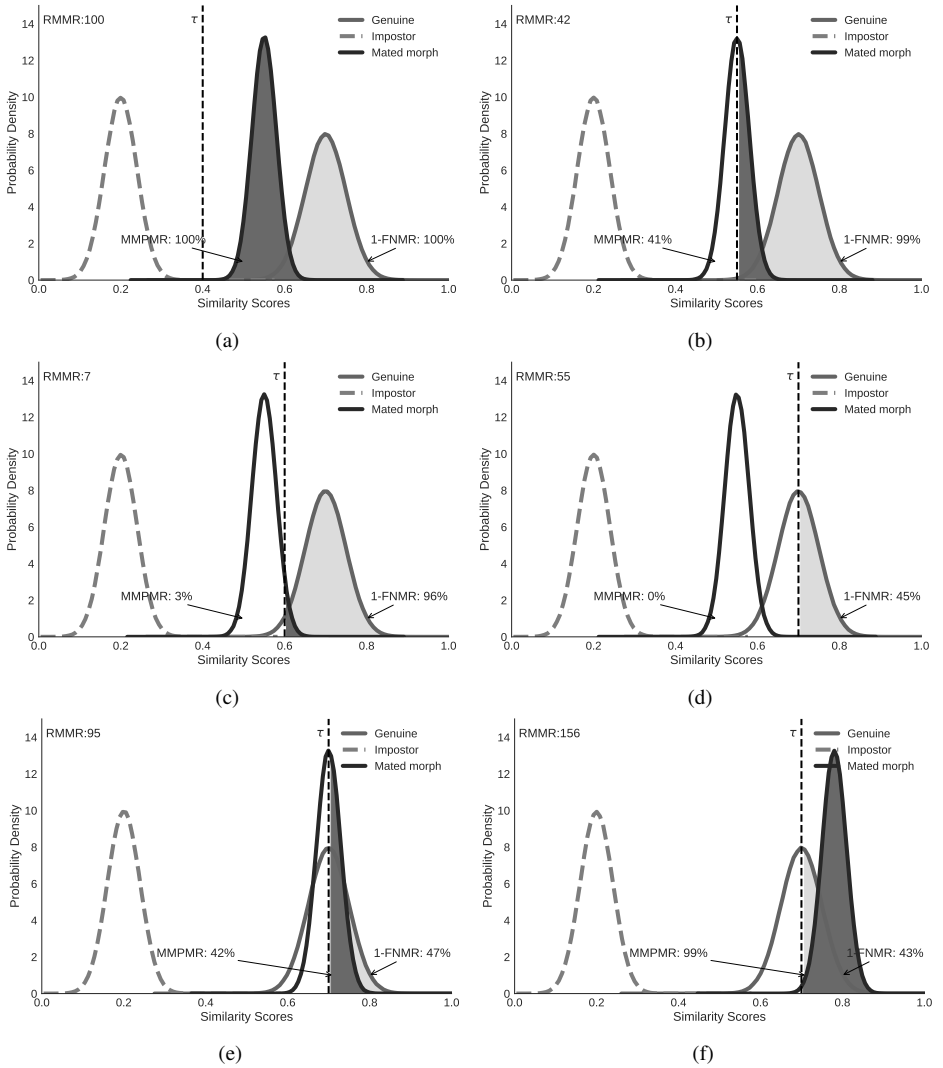


Fig. 6: Behaviour of *RMMR* for different thresholds and distributions

on a mixture of low and high quality morphs, the evaluation should be conducted separately on low and high quality morphs to ensure reliable performance measures for the different attack classes.

Finally, for morphed face image attacks in border control scenarios, the employment of comprehensible detection algorithms is strongly recommended. In order to achive justifiable and reliable results of the detection system, the system should reveal morphing-specific information, thus a clear separation of frontend (feature extractor) and backend (classifier) is needed: back-end classifiers shall be based on discriminative features subjective to morphing detection (not necessarily biometric recognition), thus a depending front-end must be employed, extracting features in a morphing-discriminant space. On assessing non morphing specific information, classifier training may be mislead regarding nuisance attributes, e.g. processing artefacts introduced by compression or scaling. In order to avoid opaque results in algorithm benchmarking, we strongly advise against algorithms, not encapsulating a clear distinction between front- and back-end, when aiming at sensitive operational real world scenarios, such as border control.

5 Conclusion

During the creation of morphing samples, multiple pitfalls have to be avoided. To that end, we have presented a summary of the observations we made so far. The key issues are the morph quality, the similarity of subjects and the consistent quality of the database. In order to allow a fair evaluation of biometric systems' vulnerability to morphing attacks, we propose new metrics, i.e. *MMPMR* and *RMMR*. Further, our experiences and considerations regarding morph detection and morph detection evaluation are summarized. The paper focuses on morphing attacks on face recognition systems, but the considerations and metrics are applicable for other modalities as well. To facilitate the use of the proposed metrics, an implementation of the evaluation metrics is provided in [Mo17].

Acknowledgment

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research, the Arts (HMWK) within CRISP (www.crisp-da.de) and the BioMobile II project (no. 518/16-30).

References

[Ab17] Abrasoft: , FantaMorph. http://www.fantamorph.com, 2017. Accessed: 2017-04-18.

[FCM17] Ferrara, M.; Cappelli, R.; Maltoni, D.: On the Feasibility of Creating Double-Identity Fingerprints. IEEE Trans. on Information Forensics and Security, 12(4), 2017.

[FFM14] Ferrara, M.; Franco, A.; Maltoni, D.: The magic passport. In: IEEE International Joint Conference on Biometrics. 2014.

[FFM16] Ferrara, M.; Franco, A.; Maltoni, D.: Face Recognition Across the Imaging Spectrum. Springer International Publishing, chapter On the Effects of Image Alterations on Face Recognition Accuracy, 2016.

[Go17] Gomez-Barrero, M.; Rathgeb, C.; Scherhag, U.; Busch, C.: Is Your Biometric System Robust to Morphing Attacks? In: Int. Workshop on Biometrics and Forensics (IWBF). IEEE, 2017.

[In05] International Organization for Standardization: Information Technology – Biometrics – Biometric Data Interchange Formats – Face Image Data. Technical Report 19794-5, JTC 1 /SC 37, Geneva, Switzerland, 2005.

[In06] International Organization for Standardization: Information technology – Biometric performance testing and reporting – Part 1: Principles and framework. ISO/IEC 19795-1:2006, JTC 1/SC 37, Geneva, Switzerland, 2006.

[In12] International Organization for Standardization: Information technology – Vocabulary – Part 37: Biometrics. ISO/IEC 2382-37:2012, JTC 1/SC 37, Geneva, Switzerland, 2012.

[In17] International Organization for Standardization: Information Technology – Biometric presentation attack detection – Part 3: Testing and reporting. ISO/IEC FDIS 30107-3:2017, JTC 1/SC 37, Geneva, Switzerland, 2017.

[KR12] Kannala, J.; Rahtu, E.: BSIF: Binarized statistical image features. 21st Int'l Conf. on Pattern Recognition (ICPR), 2012.

[MMB11] Mittal, A.; Moorthy, A. K.; Bovik, A. C.: Blind/Referenceless Image Spatial Quality Evaluator. In: 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR). 2011.

[Mo17] Morphing Vulnerability Reporting, 2017. Online available: https://github.com/dasec/mvr.

[MW02] Mansfield, A. J.; Wayman, J. L.: Best practices in testing and reporting performance of biometric devices. 2002.

[OPM02] Ojala, T.; Pietikainen, M.; Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(7), 2002.

[RB17] Rathgeb, C.; Busch, C.: On the feasibility of creating morphed Iris-Codes. In: Proc. Int. Joint Conf. on Biometrics (IJCB). pp. 1–6, 2017.

[RRB16] Raghavendra, R.; Raja, K. B.; Busch, C.: Detecting Morphed Face Images. In: 8th IEEE Int'l Conf. on Biometrics: Theory, Applications, and Systems. 2016.

[Sc17] Scherhag, U.; Raghavendra, R.; Raja, K. B.; Gomez-Barrero, M.; Rathgeb, C.; Busch, C.: On the Vulnerability of Face Recognition Systems Towards Morphed Face Attacks. In: Proc. International Workshop on Biometrics and Forensics (IWBF). 2017.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 161

What you can't see can help you – extended-range imaging for 3D-mask presentation attack detection

Sushil Bhattacharjee[1], Sébastien Marcel[2]

Abstract: High-quality custom-made 3D masks are increasing becoming a serious threat to face-recognition systems. This threat is driven, in part, by the falling cost of manufacturing such masks. Research in face presentation-attack detection (PAD) in general, and also specifically for 3D-mask based attacks, has mostly concentrated on imagery in the visible-light range of wavelengths (RGB). We look beyond imagery in the visible-light spectrum to find potentially easier solutions for the challenge of face presentation-attack detection (PAD). In particular, we explore the use of near-infrared (NIR) and thermal imagery to detect print-, replay-, and 3D-mask-attacks. This preliminary study shows that both NIR and thermal imagery can potentially simplify the task of face-PAD.

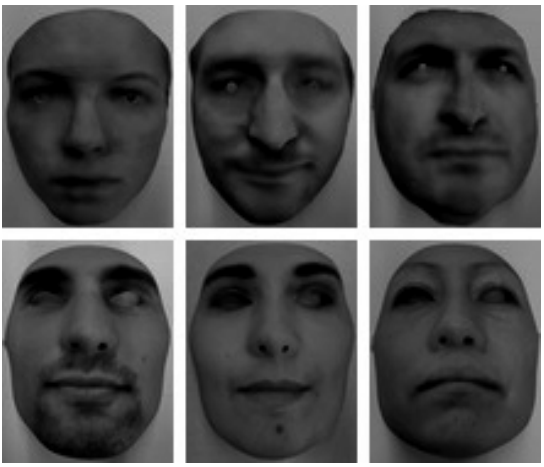
Keywords: Face Presentation Attack Detection, 3D-Masks, RGB/depth cameras, thermal cameras, NIR, LWIR.

1 Introduction

Typical face-recognition (FR) methods are highly susceptible to *presentation attacks* (PA), also commonly called *spoof attacks* [EM13, MNL14]. The term 'presentation attacks' covers both *impersonation* as well as *concealment* attacks [ISO16]. Most research efforts on presentation attack detection (PAD) so far, have considered specific kinds of attacks, and have proposed solutions to such attacks under specific conditions. Mainly 2D impersonation attacks, that is, attacks performed using 2D presentation-attack instruments (PAI) such as printed paper (*print attacks*) and replay on digital screens (*replay-attacks*) have been studied. The proposed solutions have usually considered the scenario where the biometric sensor is a single color-camera. These solutions are very well-suited for certain applications, such as for current mobile devices. The sensors on such devices are standard color cameras. Therefore, the same sensor-data is typically used for both, face-recognition, as well as for face-PAD. The PAD solutions so far have mainly relied on features extracted from color images and sophisticated machine learning algorithms, to delineate that ever-shrinking margin between the two classes (*bona fide* and attack presentations). One reason why PAD research has evolved in this direction is that adding extra sensors has often been infeasible, due to cost and space constraints.

PAs using custom-made 3D masks are receiving increasing attention [EM13, Li16]. Figure 1 shows examples of 3D-masks, including rigid masks and soft silicone masks. Highly realistic 3D-masks, such as the custom-made silicone mask shown in Fig. 1(c), are still

[1] Biometrics Group, Idiap Research Institute, Rue Marconi 19, Martigny, Switzerland. sushil.bhattacharjee@idiap.ch
[2] sebastien.marcel@idiap.ch



(a) Custom-made rigid masks

(b) Generic silicone mask

(c) Custom-made silicone mask

Fig. 1: Examples of realistic 3D masks. Approximate prices: (a) US$300, (b) US$800, and (c) US$4000.

quite expensive. However, the process of manufacturing custom-made silicone masks is evolving rapidly, and such masks will be available at accessible prices in the near future. The study presented in this paper is motivated by the threat that attacks on FR systems using high-fidelity custom masks may soon become as commonplace as print-attacks and replay-attacks today.

A new category of consumer-grade imaging devices, collectively referred to as *extended-range imaging devices*, are now available that capture data not only in visible-light wavelengths (*i.e.*, color images), but also in near-infrared (NIR) and long-wave infrared (LWIR) domains. Color/depth (RGB-D) cameras, for example, the Microsoft Kinect and Intel RealSense series of products, and thermal cameras can provide easy solutions for once challenging PAD problems.

In this article, we explore the use of such devices to simplify the task of face-PAD for both 2D and 3D attacks. This hardware based approach is especially well-suited for near-real-time scenarios such as border-control applications. To demonstrate our commitment to furthering the cause of *reproducible research*, the data and code used for the experiments presented in this paper are made freely available on the web[3].

Following a summary of related research (Section 2) on the use of extended-range imaging in face biometrics, as well as on PAD for 3D-mask attacks, we provide, in Section 3, brief descriptions of the imaging devices that have been used in this work. Experimental results are presented and discussed in Section 4. Finally, Section 5 gives a summary of this work, along with an outlook on how this work will evolve in future.

2 Related Work

Previous efforts with specialized imaging-sensors for face biometrics applications have been concentrated mainly on the problem of face-recognition. Bhowmick *et al.* [Bh11] have shown that thermal imagery can be used for face-recognition based on facial vein-patterns. Bebis *et al.* [Be06] explore the fusion of visible and thermal imagery for face-recognition applications. Although they use face-recognition performance as a metric, the main focus of their work is on evaluating different methods for fusing the data from the two modalities.

Lagorio *et al.* [La13] use 3D scans from a Vectra 3D camera in a PAD scenario, to detect curved-paper based print presentation-attacks. More recently Raghavendra *et al.* [Ra17] have published a detailed study on the vulnerability of FR systems in extended multispectral imaging domain, involving seven-band imagery covering the visible light and NIR illumination. Their work shows that all the four studied FR approaches are consistently vulnerable in all imaging-bands considered, except for the 930nm (NIR) band. This finding is consistent with previous research [Ka11] showing that the reflectance of human skin drops significantly in a narrow wavelength band centered at 970nm. Steiner *et al.* [St16] have demonstrated the use of multispectral SWIR imagery to reliably distinguish human skin from other materials, and have shown that such multispectral devices can be used for PAD.

Good quality 3D masks present clear threats in both impersonation as well as concealment categories. Erdogmus *et al.* [EM13] have published the 3DMAD dataset, using a set of custom-made rigid masks, for experiments in 3D face-PAD. This dataset has also been used by other research groups [Ag16] in other 2D face-PAD experiments. Liu *et al.* [Li16] have published the more recent HKBU-MARs dataset containing images of 3D-mask based PAs. They have proposed a remote photo-plethysmography (rPPG) based approach to detecting 3D-mask PAs. Both works ([EM13, Li16]) use several variants of local binary patterns (LBP) to demonstrate the vulnerability of FR methods to the 3D-mask PAs.

Manjani *et al.* [Ma17] present an observational study into concealment attacks using 3D-masks. They describe PAD experiments based on the SMAD dataset [Ma17], which con-

[3] https://pypi.python.org/pypi/bob.paper.BioSig2017_3DMaskPreStudy

sists of public-domain videos collected from the World-wide Web.The dataset, however, is relatively small – including only 65 genuine videos and 65 silicon-mask videos. Although observational studies such as this may indicate association between variables (in this case between the true labels of the test videos and the classifier-score), the influence of other confounding variables here cannot be ruled out. To demonstrate the efficacy of a method for detecting 3D-mask based PAs, it is important to design a controlled experiment to highlight exclusively the causal effect of 3D-masks on the resulting classifier-score.

3 Extended-Range Imagery

(a) RealSense SR300 (b) Xenics Gobi-640

Fig. 2: Cameras used in this work.

In this work we have explored the use of two different cameras – the RealSense SR300 camera from Intel, and the Gobi-640-GigE LWIR camera produced by Xenics[4] – for PAD applications.

The RealSense SR300 camera (Fig. 2(a)) is Intel's second generation depth-sensing camera. It uses a structured-light approach, based on 860nm NIR illumination, to capture depth information from a 3D surface. This camera has a relatively shallow field of view, and produces the most accurate results in the depth range of 0.2m – 1.2m. In practice, the implication for PAD applications is that the average intensity of the NIR image drops rapidly with distance. Therefore, to capture good quality images, the subject should be positioned quite close (0.3m – 0.5m) to the camera. Besides depth-information, the camera also captures color (RGB) images and NIR images. It is important to note that the two cameras (color and NIR) have different fields of view, and are not mutually calibrated. For the experiments discussed in this paper, we have used both color and NIR images at VGA resolution.

The Xenics Gobi thermal camera, shown in Fig. 2(b), covers a wavelength range of 800nm – 1200nm, and captures 16-bit images at VGA (640×480 pixels) resolution. Although the camera can take a range of lenses, we have used the standard 18mm $f/1$ lens, with a horizontal field of view of 33°.

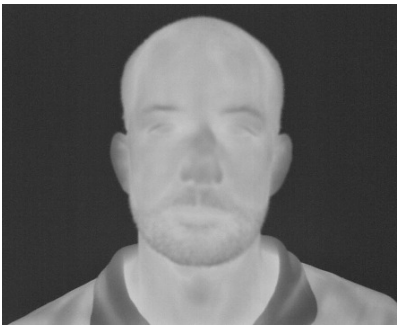
[4] Website: www.xenics.com

For both cameras, we have developed data-capture tools in-house, based on software development kits (SDK) are available for each camera. When using images outside the visible-light range of wavelengths, one practical problem that arises is that of face-detection. Most face-detection tools take a machine-learning based approach, and need to be trained with sufficient amounts of training-data. Since faces present different spectral characteristics in different wavelengths, these appearance-based face-detection schemes need to be explicitly trained for each imaging modality. We simplify the face-detection process, for each camera, by positioning the subject such that the face falls within a pre-defined rectangle (displayed on the live-display monitor). The position and size of the rectangle can be adjusted for each subject, and is recorded along with the images. Thus, for each camera, face-position information is available directly from the data-capture process.

Sample images of *bona fide* presentations are shown in Fig. 3. The images in Fig. 3(a) and (b) have been captured using the SR300 RGB-D camera, and show the *bona fide* presentation in visible wavelengths (RGB) and NIR band respectively. Figure 3(c) shows an image captured using the Xenics Gobi thermal camera, illustrating the appearance of a *bona fide* presentation in the LWIR band.

(a) Visible (RGB)

(b) NIR



(c) Thermal (LWIR)

Fig. 3: Examples of *bona fide* presentation images, as seen in different wavelength bands. Images in (a) visible (RGB) and (b) NIR wavelength-bands have been captured using the SR300. The LWIR band image (c) has been captured using the Xenics Gobi camera.

4 Experiments

In this section we present experimental results for PAD based on extended-range imagery. Based on some initial tests, we concluded that NIR images have the potential to easily detect 2D and 3D PAs. In the experiments reported here, we have specifically investigated the following questions. (1) Can NIR imagery be useful in detecting 2D PAs? (2) Is it possible to detect 3D-mask attacks (of both rigid and flexible varieties) in NIR images? (3) Can we use thermal (LWIR) images to detect custom-made flexible-mask PAs?

4.1 PAD Using NIR Imagery

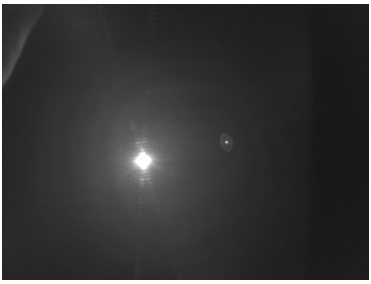
We start by confirming the intuition that NIR imagery can simplify the task of detecting 2D PAs such as print- and replay-attacks. Figure 4 illustrates typical images captured by the SR300 camera for various kinds of PAs. Fig. 4(a) and (d) show the appearance of a print-attack in color- and NIR-imagery, respectively. We note that although the face may be detected in the color image (Fig. 4(a)), even a simple image-histogram analysis would be sufficient to determine that no face present in the corresponding NIR image (Fig. 4(d)). The analysis for the digital-replay attack shown in Fig. 4(b) and (e) is analogous.

Fig. 4(c) and (f) show the color and NIR images for the same rigid mask attack. Visual inspection of the mask region shows that the simulated facial features, such as the painted eye-brows and moustache, are entirely suppressed in the NIR image, and the mask region has a texture-less appearance. Intuition tells us that detecting such surfaces in NIR images should be reasonably easy. Contrary to initial expectations, however, detecting such 3D mask PAs in NIR images is not straightforward. Comparing Fig. 4(f) with Fig. 3(b), we see that the NIR image presents similar image characteristics for both *bona fide* and 3D-mask attack presentations. Indeed, our preliminary tests with NIR-860nm band images showed that lower-order statistics (intensity-histograms, histograms of oriented gradients (HOG) and gray-level co-occurence matrices (GLCM)) of the 3D-mask presentations are quite similar to those of *bona fide* presentations. Detection of 3D mask PAs in NIR images would require more complex processing such as, modeling of the peri-occular region.

Although counter-intuitive, this result is quite logical. The NIR wavelength used by the RGB-D cameras has been deliberately chosen to be such that the illumination is strongly reflected by most kinds of surfaces, including human-skin. This is imperative for the primary purpose of the camera – capturing depth-information using structured-NIR illumination.

(a) Print; Color

(b) Print; NIR



(c) iPad; NIR

(d) iPad; Color

(e) 3D Mask; Color

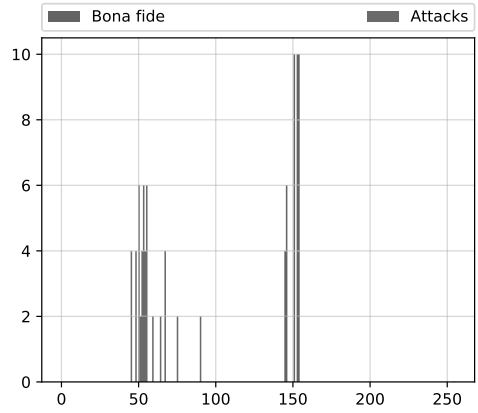
(f) 3D Mask; NIR

Fig. 4: Various PAs, as seen by the RGB-D (SR300) camera. Three kinds of attacks – print, replay, and 3D-mask – are shown. Left column (a), (c), (e): the appearance of print, replay, and 3D-mask attack, respectively, in visible color wavelengths. Right column (b), (d), (f): corresponding images of the respective scenes under NIR illumination.

4.2 PAD Using Thermal Imagery

(a) *Bona fide*

(b) 3D-mask attack



(c) Intensity distribution

Fig. 5: Examples of thermal images of (a) *bona fide* and (b) 3D-mask attack presentations. (c) Histograms of average-intensity over the face-region, for mask- and *bona fide* presentations, computed over a small dataset of thermal images.

Thermal (LWIR) images are very well suited for detecting 3D-masks PAs with high certainty. Figure 5 shows images from the Xenics Gobi thermal camera. The mask in Fig. 5(b), being cooler than the body-temperature of the subject, is clearly demarcated. Figure 5(c) shows distributions of the average pixel-intensity of a small region centered on the face. This plot has been generated based on a small dataset of *bona fide* and 3D-mask attack presentations using five subjects and six rigid masks. The intensity distribution for the rigid-mask presentations is plotted in red, and the distribution for the *bona fide* presentations is plotted in green. Although this plot is based on a very small dataset, it indicates that the average intensity over the face-region is significantly lower when a rigid-mask is used.

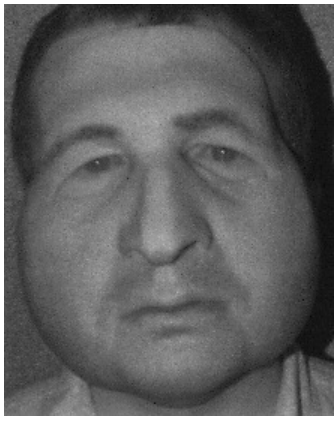
Flexible silicone masks are often hand-finished, and offer greater color and texture fidelity, and therefore, pose a greater threat, compared to rigid masks. The images in Fig. 6 allow us

to compare the appearances of *bona fide* presentations and flexible custom-made silicone-mask attack presentations, under visible-light and NIR illuminations. The mask shown here (Fig. 6(c), (d)), which is the same as the mask shown in Fig. 1(c), has been custom-designed to match the face of the subject shown in Fig. 6(a) and (b)[5]. Although the mask may be easily apparent to the human observer, FR systems are quite vulnerable to such mask-attacks.

(a) *Bona fide*; RGB

(b) *Bona fide*; NIR

(c) Flex. mask; RGB



(d) Flexible mask; NIR

Fig. 6: Comparison of *bona fide* and flexible silicone mask attack presentations in visible-light and NIR illuminations. Images captured using the SR300 camera.

Table 1 shows the face-recognition scores for the various attack-attempts on the reference-subject shown in Fig. 6(a)&(c). The pre-trained VGG-Face neural network [Pa15] has been used for face-recognition in this experiment. Specifically, the table shows scores for

[5] The mask was manufactured based on 3D facial scans of the subject, using the Intel RealSense F200 camera, which is the predecessor of the SR300 camera. Additional color photographs were used to ensure high-quality finish for the visual appearance of the mask.

five zero-effort impostor (ZEI) presentations, one genuine presentation, and one attack presentation made using a custom-made silicone-mask. Scores are shown for two image-modalities – color images and NIR images. The first row (labeled 'RGB') shows scores of the various presentations using color images, where the target identity is the image in Fig. 6(a). Scores for a similar experiment using NIR images, with the image in Fig. 6(b) as the target identity, are shown in the last row of the table. In both experiments we can see that the score for the genuine presentation is at least an order of magnitude higher than the scores for the ZEI presentations (Subjects 1-5). It is interesting to note that, in both illumination wavelength bands, the score for the custom-mask attack is much closer to that of the genuine presentation, than to the ZEI presentations. This small-scale experiment cannot be attributed any statistical significance. It does, however, emphasize the necessity for a large-scale study involving attacks with high-quality custom-made silicone masks.

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Genuine	Mask-attack
RGB	-0.344	-0.253	-0.265	-0.293	-0.263	-0.038	-0.163
NIR	-0.330	-0.227	-0.280	-0.216	-0.198	-0.028	-0.096

Tab. 1: Face recognition scores using VGG network.

4.3 Discussion

Fig. 7: Silicone masks can warm up very quickly when worn.

Preliminary experimental results discussed above show that extended-range imaging devices can drastically simplify the task of face-PAD. The task of face-PAD using such devices, however, in not entirely straightforward. For example, the main reason why the face in the print-attack is not visible in 4(d)) is that the ink used here is not IR-reflective[6]. There are, however, IR-reflective inks available on the market, that will show a strong response in NIR-band images [Ch15]. Further experimentation is therefore necessary to tackle the challenge print-attacks constructed using IR-reflective inks.

In Fig. 5 we have seen how rigid 3D-mask attacks can be easily detected using thermal cameras, because the masks have a much lower temperature than average human body temperature. This advantage is lost to some extent, when dealing with flexible silicone masks, which can warm up very quickly when in contact with warmer objects, such as an attacker's face. Figure 7 illustrates the evolution of the temperature of a silicone mask when worn by a human subject. The figure shows six frames of a time-lapse sequence (captured using the Xenics Gobi camera), with an interval of 30 seconds between frames.

[6] Note that the specific print-attack used in this example has been printed according to the specifications prescribed in the Norwegian project SWAN.

From left to right, the change in temperature of the mask is evident, especially in the forehead region, where the mask makes good contact with the subject's face. The top of the mask can be clearly distinguished in the first frame on the left. In the right-most frame, this region seems to be almost as warm as the subject.

There is a clear need for research into PAD of flexible 3D masks, where the limits of thermal imagery for detecting custom-made masks is explored under various conditions, using a large set of silicone masks. Thermal cameras such as the Xenics Gobi used in this work are still quite expensive. However, low-cost options are now becoming increasingly available. Cameras such as the FLIR-One and the Seek-Thermal Compact Pro are designed to work with mobile phones, offer adequate spatial resolution, and are available at relatively modest prices.

5 Conclusions

Due to the affordability of modern RGB-D cameras, researchers are beginning to look into such cameras for face-recognition applications. These cameras typically use NIR illumination to recover depth information from the scene. Some RGB-D cameras also return 2D videos in visible-light and NIR wavelengths. In this paper we have presented a preliminary study into the utility of NIR and LWIR imagery for face-PAD. Our tests with Intel's RealSense SR-300 camera show that images in NIR wavelength-band can be used to easily detect various 2D presentation-attacks, such replay-attacks and certain kinds of print-attacks. Some kinds of printer-ink do show a strong response under NIR illumination at certain wavelengths. In future work we will investigate methods for detecting print-attacks created using such inks. Contrary to initial expectations, however, monochromatic NIR imagery, of the kind provided by low-cost RGB-D cameras, may not be effective for straightforward detection of 3D-mask attacks.

Realistic custom-made silicone masks will soon be available at affordable prices. Tt is, therefore, imperative to develop face-PAD methods that are robust to 3D-mask based attacks. The examples presented here indicate that the use of NIR and LWIR imagery for detecting rigid as well as flexible 3D masks seem to be promising research directions. We would like to invite the entire biometrics research community to pool resources to create and share a large and diverse data-set for the purposes of such research.

Thermal cameras have been available for several decades. Until recently, however, they have been very expensive, and have not been considered for PAD applications for reasons of cost. Low-cost thermal cameras, such as the FLIR-One[7] and the Seek-Thermal Compact Pro[8], have recently appeared on the market. These devices are designed to work with mobile phones. They offer adequate spatial resolution, and are available at reasonable prices. In future experiments we plan to explore the applicability of such cameras for PAD.

[7] Website: www.flir.com/flirone
[8] Website: www.thermal.com/compact-series.html

Acknowledgement

This work has been supported by the EU H2020 project TeSLA, the Norwegian project SWAN, and the Swiss Center for Biometric Research and Testing. We gratefully acknowledge the critical help from our colleague, Mr. Guillaume Clivaz, who implemented the image-capture applications for the two cameras used in this work.

References

[Ag16] Agarwal, A. et al.: Face Anti-Spooïng using Haralick Features. In: Proc. IEEE Intl. Conf. BTAS. Niagara Falls, NY, USA, 2016.

[Be06] Bebis, G. et al.: Face Recognition by Fusing Thermal Infrared and Visible Imagery. Image and Vision Computing, 24(7):727–742, July 2006.

[Bh11] Bhowmick, M et al.: Thermal Infrared Face Recognition - a Biometric Identification Technique for Robust Security System. In (Corcoran, P., ed.): Reviews, Refinements and New Ideas in Face Recognition. Intech, Rijeka, Croatia, pp. 135–162, 2011.

[Ch15] Chingovska, I. et al.: Face recognition in extended imaging domain. In (Bourlai, T., ed.): Face Recognition Across the Imaging Spectrum. Springer, pp. 165–194, 2015.

[EM13] Erdogmus, N.; Marcel, S.: Spooïng in 2D Face Recognition with 3DMasks and Anti-spooïng with Kinect. In: Proc. IEEE Intl. Conf. BTAS. Washington D.C., 2013.

[ISO16] : ISO/IEC DIS 30107-1. Information Technology – Biometric Presentation Attack Detection – Part 1: Framework. Iso standard, Geneva, CH, January 2016.

[Ka11] Kanzawa, Y. et al.: Human Skin Detection by Visible and Near-Infrared Imaging. In: Proc. IAPR Conf. on Machine Vision Applications (MVA2011) . Nara, Japan, 2011.

[La13] Lagorio, A. et al.: Liveness Detection based on 3D Face Shape Analysis. In: Proc. IEEE Intl Workshop on Biometrics and Forensics (IWBF). Lisbon, Portugal, 2013.

[Li16] Liu, S. et al.: A 3D Mask Face Anti-spoofing Database with Real World Variations. In: Proc. IEEE Conf. on Comp. Vision and Patt. Rec. Workshop (CVPRW). Las Vegas, 2016.

[Ma17] Manjani, I. et al.: Detecting Silicone Mask-Based Presentation Attack via Deep Dictionary Learning. IEEE Trans. Info. Forensics and Security, 12(7):1713 – 1723, 2017.

[MNL14] Marcel, S.; Nixon, M. S.; Li, S. Z., eds. Handbook of Biometric Anti-Spoofing. Springer-Verlag, 2014.

[Pa15] Parkhi, Omkar M. et al.: Deep Face Recognition. In: British Machine Vision Conference. 2015.

[Ra17] Raghavendra, R. et al.: On the Vulnerability of Extended Multispectral Face Recognition Systems Towards Presentation Attacks. In: Proc. IEEE Intl. Conf. on Identity, Security and Behavior Analysis (ISBA). New Delhi, 2017.

[St16] Steiner, H. et al.: Design of an Active Multispectral SWIR Camera System for Skin Detection and Face Verification. Journal of Sensors, (1):1 – 8, 2016. Article ID 9682453, Special Issue on Multispectral, Hyperspectral, and Polarimetric Imaging Technology.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 173

Impact of bandwidth and channel variation on presentation attack detection for speaker verification

Héctor Delgado[1], Massimiliano Todisco[1], Nicholas Evans[1], Md Sahidullah[2], Wei Ming Liu[3], Federico Alegre[3], Tomi Kinnunen[2], Benoit Fauve[3]

Abstract: Vulnerabilities to presentation attacks can undermine confidence in automatic speaker verification (ASV) technology. While efforts to develop countermeasures, known as presentation attack detection (PAD) systems, are now under way, the majority of past work has been performed with high-quality speech data. Many practical ASV applications are narrowband and encompass various coding and other channel effects. PAD performance is largely untested in such scenarios. This paper reports an assessment of the impact of bandwidth and channel variation on PAD performance. Assessments using two current PAD solutions and two standard databases show that they provoke significant degradations in performance. Encouragingly, relative performance improvements of 98% can nonetheless be achieved through feature optimisation. This performance gain is achieved by optimising the spectro-temporal decomposition in the feature extraction process to compensate for narrowband speech. However, compensating for channel variation is considerably more challenging.

Keywords: presentation attack detection, speaker verification, bandwidth and channel variation

1 Introduction

While automatic speaker verification (ASV) [RR95, KL10, HH15] offers a convenient, reliable and cost-effective approach to person authentication, vulnerabilities to presentation attacks [IS16], previously referred to as spoofing, can undermine confidence and form a barrier to exploitation. By masquerading as enrolled clients, fraudsters can mount attacks to gain unauthorised access to systems or services protected by biometrics technology.

Presentation attacks in the context of ASV can be performed with impersonation, speech synthesis, voice conversion and replay [EKY13]. While the study of impersonation has received attention, e.g. [Ha15], replay, speech synthesis and voice conversion are assumed to pose the greatest threat [Wu15]. Speech synthesis and voice conversion presentation attacks combine suitable training or adaptation data with sophisticated algorithms which generate voice samples whose spectral characteristics resemble those of a given target speaker. In contrast, replay spoofing attacks require neither specialist expertise nor equipment and can hence be mounted by the lay person with relative ease. Replay attacks involve the re-presentation to an ASV system of another person's speech which is captured beforehand, possibly surreptitiously, for instance during an access attempt.

The study of presentation attack detection (PAD) for ASV is now an established area of research [Wu15]. The first competitive evaluation, namely the ASV spoofing and counter-

[1] EURECOM, Department of Digintal Security, France, {delgado,todisco,evans}@eurecom.fr
[2] University of Eastern Finland, School of Computing, Finland, {sahid,tkinnu}@cs.uef.fi
[3] Validsoft Ltd., United Kingdom, {jasmin.liu,federico.alegre,benoit.fauve}@validsoft.com

measures (ASVspoof) challenge [Wu17], was held in 2015. It promoted the development PAD solutions to protect ASV from voice conversion and speech synthesis attacks.

Since the first ASVspoof 2015 evaluation, the community has started to consider a number of more practical aspects of PAD. Some recent work has explored the impact of additive noise on reliability [Ha16, Ti16] and the benefit of speech enhancement and multi-condition training as a means of improving robustness [Ha16, Yu16].

Other likely influences on PAD performance, e.g. bandwidth and channel variability, have received comparatively little attention to date [GGS15, VL11]. Given the prevalence of ASV technology in telephony applications were bandwidth is typically low and where coding, packet loss and other non-linear effects have potential to degrade performance, these aspect require attention. However, the ASVspoof 2015 database contains high quality, high bandwidth recordings. The RedDots Replayed database [Ki17c] which was generated from the text-dependent ASV RedDots database [Le15], was introduced recently to support the development PAD solutions for replay presentation attacks. While exhibiting variation in terms of recording devices and environmental conditions, and hence representing a greater degree of practical, real-life variability, it still contains wideband audio (16kHz).

The work reported in this paper has accordingly sought to investigate the impact of bandwidth and channel variation on PAD reliability for ASV. The work was performed with bandwidth-limited and coded versions of the ASVspoof 2015 and RedDots Replayed databases (covering 3 different types of presentation attacks, namely speech synthesis, voice conversion and replay), generated through band-pass filtering, downsampling and coding. The work was performed with two PAD solutions, namely linear frequency cepstral coefficients [SKH15] and constant Q cepstral coefficients [TDE16, TDE17], both of which achieve competitive performance for the ASVspoof 2015 database with a relatively simple back-end classifier. It is stressed that the objective of the work reported here is to assess the impact on PAD reliability of bandwidth and channel variation. While an issue of undoubtable importance, the work is NOT concerned with generalisation.

2 Presentation attack databases

The work reported in this paper was performed using two publicly available databases.

2.1 ASVspoof 2015

The ASVspoof initiative [Wu17] was the first community-led effort to collect a common database to support research in spoofing and countermeasures. The ASVspoof 2015 database contains a mix of *bona fide* (genuine speech without attack) and spoofed speech. All bona fide speech data is sampled at 16kHz and was recorded in a semi-anechoic chamber with a solid floor [Wu17]. Spoofed speech is generated with 10 different speech synthesis and voice conversion algorithms. In order to promote generalised PAD systems, only 5 of these were used to generate training and development subsets whereas an evaluation subset was generated with the full 10. In this paper, the development set containing genuine and spoofed speech using 5 different attacks is used. Table 1 shows database statistics. Full details of the ASVspoof 2015 database and example PAD results are available in [Wu17].

Tab. 1: *Statistics of the ASVspoof 2015 database: number of speakers (M=male, F=female), and number of genuine and spoofed trials.*

Partition	#Speakers (M / F)	#Genuine trials	#Spoofed trials
Training	10 / 15	3750	12625
Development	15 / 20	3497	49875

2.2 RedDots Replayed

The **RedDots Replayed** database [Ki17c] was designed to support the development of PAD solutions for replay attacks in diverse recording and playback environments. RedDots Replayed is based upon the re-recording of the original RedDots database [Le15] (part 01, male speakers) which contains speech data comprising 10 common passphrases recorded in a number of acoustic conditions using mobile devices with a sampling rate of 16kHz. Replayed speech is generated with one of 16 different recording devices, 15 different playback devices and various different acoustic conditions, including both controlled and more variable (unpredictable) conditions. Controlled condition recordings are made in a quiet office/room whereas variable condition recordings are made in noisier environments. A training subset contains only controlled condition recordings whereas an evaluation subset contains both controlled and variable condition recordings. Table 2 shows database statistics. Full details of the RedDots replayed database and example presentation attack detection results are available in [Ki17c]. A subset of the RedDots Replayed database is also used in the ASVspoof 2017 challenge[4] data [Ki17a, Ki17b].

Tab. 2: *Statistics of the RedDots Replayed database: number of speakers (male), and number of genuine and spoofed trials.*

Partition	#Speakers	#Genuine trials	#Spoofed trials
Training	10	1508	9232
Development	39	2346	16067

2.3 Bandwidth reduction and channel simulation

PAD performance was assessed with different versions of each database: (i) the original full-band versions; (ii) bandwidth-reduced versions, and (iii) versions with additional channel variation simulated with the Idiap acoustic simulator software[5].

Bandwidth reduction involves downsampling from 16kHz to 8kHz. ITU G.151[6] compliant bandpass filtering is applied with a gain of -3dB at the passband edges of 300Hz and 3400Hz. The original and bandwidth-reduced versions are referred to from hereon as **wideband** (WB) and **narrowband** (NB).

[4] `http://www.asvspoof.org/`
[5] `http://github.com/idiap/acoustic-simulator`
[6] `https://www.itu.int/rec/T-REC-G.151-198811-W/en`, accessed: 2017-08-07

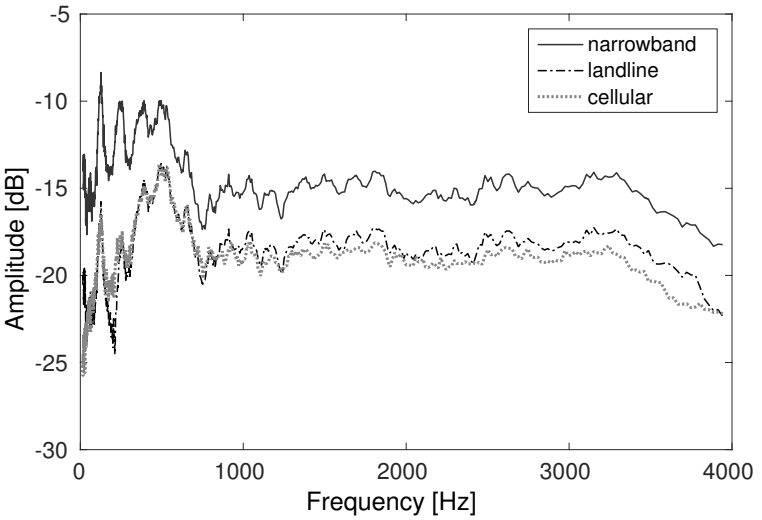


Fig. 1: Average long-term spectra for the utterance 'He's worked for several years in the United States' for narrowband, landline and cellular channels.

Codec simulations employ a common ITU G.712[7] compliant bandpass filter. This is combined with a-law coding[8] at a rate of 64kbit/s for landline telephony and with an adaptive multi-rate narrowband (AMR-NB) codec[9] at a rate of 7kbit/s for cellular telephony. These two scenarios are referred to as **landline** (L) and **cellular** (C), respectively. Figure 1 illustrates the distortion in the long-term average spectrum for landline and cellular coded signals compared to the original narrowband signal for an arbitrary speech utterance from the ASVspoof 2015 database. These spectra were obtained with the constant Q transform (CQT, see Section 3). In addition to broad attenuation, the plots illustrates substantial spectral distortion, especially at lower and higher frequencies. The distortion is particularly pronounced for the cellular-coded signal.

3 Presentation attack detection

The work was performed with two different PAD systems. A backend Gaussian mixture model (GMM) classifier with two classes, one for fona fide speech and one for spoofed speech is common to both systems. Models are learned using bona fide and spoofed data from their respective training subsets and with an expectation maximisation algorithm. According to independent results, e.g. [PP15, SKH15, TDE17], such a simple classifier often provides competitive or even better performance compared to other, more sophisticated algorithms. The score for a given trial is computed as the log-likelihood ratio of the test speech sample given the two GMMs for bona fide and spoofed speech. The frontends are described below. Neither employs voice activity detection.

[7] `https://www.itu.int/rec/T-REC-G.712/en`, accessed: 2017-08-07
[8] `https://www.itu.int/rec/T-REC-G.711-198811-I/en`, accessed: 2017-08-07
[9] `https://www.itu.int/rec/T-REC-G.711-198811-I/en`, accessed: 2017-08-07

The **linear-frequency cepstral coefficient** frontend is the best performing system from [SKH15]. The energy outputs of a uniformly-spaced, triangular filterbank are processed by the discrete cosine transform (DCT) to derive cepstral coefficients using an analysis window of 20ms with a 10ms shift. Since LFCC features are computed with linearly-spaced filters, the frequency resolution is explicitly related to the number filters. Increasing the number improves the frequency resolution and captures more detailed spectral characteristics. While the original work [SKH15] used 20 filters, use of a greater number was found to improve performance. For work reported here, the number of filters is optimised first for WB and then for NB data.

Constant Q cepstral coefficients. The second front-end involves constant Q cepstral coefficients (CQCCs) [TDE16, TDE17] which combine the constant Q transform (CQT) [YB78] with standard cepstral analysis. In contrast to Fourier techniques, the centre/bin frequencies of the CQT scale are geometrically distributed [RB79]. The centre frequency f_k for the k-th frequency bin is given by

$$f_k = f_{min} 2^{\frac{k-1}{B}} \tag{1}$$

where f_{min} is the minimum frequency considered and B is the number of bins per octave. Higher values of B provide greater frequency resolution but reduced time resolution, while lower values of B provide greater time resolution but smaller frequency resolution. B thus determines the trade-off between frequency and time resolutions and is a major optimisation parameter for CQT-based analysis. Note that the CQCC analysis window length and shift is effectively variable in order to maintain a constant Q factor (trade-off between centre frequency and filter width) across frequency bins. Full details of CQCC extraction are described in [TDE17].

4 Experimental work

This section reports an assessment of bandwidth and channel variation impacts on PAD performance. All experiments were performed with the standard protocols in [Ki17c, Wu17] (see Section 2). Assessments are based on the threshold-free equal error rate (EER_{pad}) metric for a bona fide/presentation attack discrimination task. EER_{pad} is the operating point where the attack presentation classification error rate, APCER (equivalent to the false alarm rate, FAR, in binary classification tasks), and the bona-fide presentation classification error rate, BPCER (equivalent to miss rate in binary classification tasks), are equal. Shown first are baseline experiments using the original high-quality WB versions of the ASVspoof 2015 development set (in the following referred to as ASVspoof) and RedDots Replayed database. The use of the ASVspoof development set alone avoids any influence of results on presentation attacks for which no training data is available; **this paper is not concerned with generalisation aspects**. Then, the adopted methodology is summarised as follows:

- Baseline experiments using the original high-quality WB databases were performed.
- Identical experiments using NB versions of the same databases were performed to evaluate performance for bandwidth-reduced audio.

- Feature extraction configurations are optimised to improve performance for bandwidth-reduced audio.
- A final set of experiments evaluate the robustness of optimised PAD solutions in the face of additional speech coding.

4.1 Wideband baseline

Baseline results for LFCC and CQCC features and the original WB databases (no down-sampling nor channel simulation) are presented in Table 3 ("Wideband" rows). LFCCs include 20 delta (D) and 20 acceleration (A) coefficients [SKH15] computed using 30 filters while CQCCs include 20 A coefficients [TDE17]. These configurations were optimised for the ASVspoof database. Error rates for LFCC features are twice those of CQCC features. Error rates for the RedDots Replayed database are markedly higher than for the ASVspoof database, albeit that these results were generated using un-optimised feature configurations.

Tab. 3: *Performance of LFCC and CQCC PAD systems in terms of* EER_{pad} *(%) for ASVspoof development and RedDots Replayed databases for* ***WB*** *and* ***NB*** *data. PAD systems were not optimized for NB data. S=static, D=delta, A=acceleration.*

	Feature		ASVspoof 2015	RedDots Replayed
Wideband (16 kHz)	LFCC	DA	0.11	6.18
	CQCC	A	0.05	3.27
Narrowband (8 kHz)	LFCC	S	6.60	13.30
		D	3.38	9.02
		A	4.06	8.24
		SD	3.72	10.27
		SA	3.17	9.56
		DA	**1.64**	**8.12**
		SDA	2.27	8.59
	CQCC	S	10.39	7.13
		D	10.93	3.18
		A	**9.92**	**2.07**
		SD	5.64	4.05
		SA	5.90	4.18
		DA	8.97	2.14
		SDA	5.71	2.88

4.2 Bandwidth reduction

Table 3 ("Narrowband" rows) shows results for the NB versions of ASVspoof and RedDots Replayed databases. Results are shown for both LFCC and CQCC features using different combinations of static (S), delta (D) and acceleration (A) coefficients. Results in Table 3 show that, for the ASVspoof database, performance is significantly degraded for

both LFCC and CQCC features. For LFCC features, the EER_{pad} increases from 0.11% to 1.64% whereas that for CQCC features increases from 0.05% to 9.92%. In addition, for CQCC, SD configuration further reduces the error rate of A configuration further down to 5.64%.

For the RedDots Replayed database, performance for LFCC features degrades from 6.18% to 8.12%. For CQCC features, results improve, with the EER_{pad} dropping from 3.27% to 2.07%. Our analysis suggests that this is because the salient information for replay detection is contained within low frequencies for which CQCC features have better resolution. The same behaviour is not observed for LFCC features, however. This is because LFCC features may lack sufficient resolution at low frequencies to capture the same information captured by CQCC features.

While it is not entirely surprising that different features are best for the ASVspoof and RedDots Replayed databases – they contain presentation attacks of a different nature – performance is sensitive to the particular configuration. Whereas DA and A combinations give the best performance for WB ASVspoof data for LFCC and CQCC features respectively, DA and SD combinations give the best performance for NB data. Performance for the RedDots Replayed database is more consistent with DA and A configurations again giving the best performance.

4.3 Feature optimisation

Tab. 4: *Optimisation of number of filters for **LFCC** features for **NB** ASVspoof development and RedDots databases in terms of* EER_{pad} *(%) for different configurations of static (S), delta (D) and acceleration (A) coefficients.*

		20	30	40	50	60	70	80
ASVspoof 2015	S	**5.74**	6.60	6.19	6.12	6.34	6.45	6.52
	D	4.48	3.38	3.28	**3.19**	3.21	3.21	3.25
	A	5.21	4.06	4.05	4.05	3.94	**3.91**	4.04
	SD	**3.48**	3.72	3.62	3.67	3.64	3.65	3.49
	SA	3.27	3.17	**3.04**	3.21	3.13	3.16	3.08
	DA	2.10	1.64	1.67	1.49	1.50	**1.44**	1.55
	SDA	2.34	2.27	2.18	2.21	2.13	2.16	**2.06**
RedDots Replayed	S	13.71	**13.30**	13.51	13.97	14.45	15.18	15.30
	D	9.06	**9.02**	9.51	9.66	10.14	10.05	10.60
	A	**8.13**	8.24	8.48	8.52	8.97	9.15	9.26
	SD	10.67	**10.27**	10.87	11.64	11.61	11.72	11.74
	SA	9.97	**9.56**	10.14	10.38	10.72	11.08	11.13
	DA	8.40	**8.12**	8.40	9.08	8.72	9.04	9.40
	SDA	9.11	**8.59**	9.63	9.65	10.17	10.57	10.53

Reported now are results for optimised LFCC and CQCC features for NB data. For LFCC features, optimisation is performed by varying the number of filters. The dimensionality of static features is fixed by considering first 20 coefficients after the DCT. Table 4 reports results for ASVspoof and RedDots Replayed databases where the number of filters is

varied between 20 and 80. For the ASVspoof database, performance is improved for a higher number of filters. The best performance is obtained with 70 filters and dynamic coefficients (DA). However, for the RedDots Replayed database, the optimal number of filters is 30 while performance degrades for higher numbers.

Tab. 5: *Optimisation of the number of frequency bins per octave B for **CQCC** features for **NB** ASVspoof and RedDots Replayed databases in terms of* EER_{pad} *(%) for different configurations of static (S), delta (D), and acceleration (A) coefficients.*

	B	192	96	48	24	12	6
ASVspoof 2015	S	17.23	10.39	5.25	2.95	**1.93**	3.06
	D	16.01	10.93	7.11	5.64	**4.53**	6.27
	A	14.73	9.92	8.08	6.40	**4.88**	8.69
	SD	10.97	5.64	2.72	1.00	**0.28**	0.37
	SA	10.45	5.90	3.35	1.05	**0.17**	0.31
	DA	13.29	8.97	6.25	4.44	**3.54**	5.70
	SDA	10.30	5.71	2.60	0.84	**0.16**	0.27
RedDots Replayed	S	**6.57**	7.13	8.82	10.06	9.68	
	D	3.50	**3.18**	3.46	7.55	11.68	
	A	2.50	**2.07**	3.20	4.65	9.21	
	SD	**3.88**	4.05	5.43	7.20	7.74	-
	SA	**3.85**	4.18	5.63	7.30	8.15	
	DA	2.73	**2.14**	2.6	4.82	11.05	
	SDA	**2.86**	2.88	3.86	6.22	8.44	

Table 5 shows optimisation results for CQCC features. Performance is illustrated for different combinations of S, D and A coefficients and as a function of the number of bins per octave B involved in the CQT computation. The combination of SDA coefficients gives the best performance for the ASVspoof database (0.16% EER_{pad} for B=12) whereas A coefficients alone give the more consistent performance for the RedDots database (2.07% EER_{pad} for B=96). In terms of general trends, smaller values of B give better performance for the ASVspoof database whereas larger values of B give better performance for the RedDots database. This would suggest that the detection of voice conversion and speech synthesis attacks requires a spectro-temporal analysis with higher time resolution. Conversely, the reliable detection of replay attacks requires a higher frequency resolution.

4.4 Channel simulation

For experiments described above, PAD algorithms were optimised for a 'generic' telephony scenario through the downsampling of original WB data to NB data. Experiments reported here focus on the evaluation of PAD systems on more challenging data with simulated landline (L) and cellular (C) channel variation. Results are presented in Table 6 for the optimised PAD systems corresponding to Tables 4 and 5. LFCC features have dynamic coefficients (DA) computed using 70 filters for the ASVspoof database. For the RedDots Replayed database, features are the same, except for 30 filters. Performance degrades significantly for both landline and cellular scenarios, more so for the ASVspoof database than for the RedDots Replayed database.

Tab. 6: *Performance of optimum configurations found in Section 4.3 applied to the ASVspoof and RedDots Replayed databases with* ***simulated cellular (C) and landline (L) channels*** *(results for narrowband (NB) also included for comparison).*

	ASVspoof			RedDots Replayed		
	NB	L	C	NB	L	C
LFCC	1.44	6.05	**11.09**	8.12	8.38	**10.14**
CQCC	**0.16**	**1.86**	12.96	**2.07**	**3.10**	12.32

CQCC features involve the full SDA configuration with B=16 frequency bins per octave for the ASVspoof database and A coefficients with B=96 frequency bins per octave for the RedDots Replayed database. Performance again degrades significantly for both landline and cellular scenarios and, again, much more for the latter. The relative degradation for CQCC features in the case of the cellular scenario is significantly greater than for LFCC features. This could indicate that, despite seemingly better performance for matched conditions, CQCC features are more sensitive to channel variation than LFCC features. Given that both landline and cellular scenarios share the same bandpass filtering, the degradation stems from the use of different codecs. The AMR-NB codec has a high compression rate of 7kbits/s. This degradation in performance most likely stems from aggressive compression and the consequential loss of frequency components which are crucial for presentation attack detection.

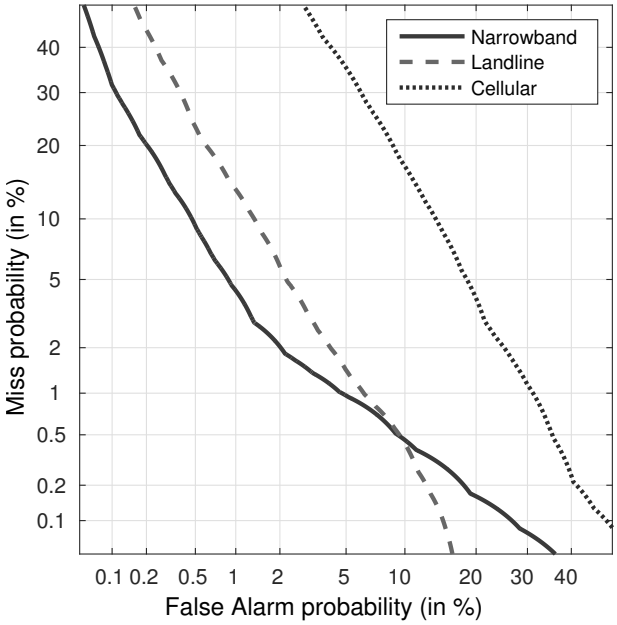


Fig. 2: DET plots for narrowband, landline and cellular scenarios on the RedDots replayed database.

To further illustrate PAD performance degradation due to codec effects, Figure 2 shows DET plots of the CQCC PAD system for generic narrowband, landline and cellular scenarios on the RedDots replayed database (replay attacks). PAD on narrowband data is more

accurate than on landline data for a wide range of operation points. PAD performance on cellular data is importantly degraded for the complete range of operation points.

5 Conclusions

This paper reports an investigation of bandwidth and channel variation on the reliability of presentation attack detection (PAD) for automatic speaker verification. Experiments were performed using two common databases of spoofed speech, namely ASVspoof 2015 and RedDots Replayed which, together, contain a variety of different presentation attacks. Results show that the performance of two state-of-the-art PAD solutions optimised for WB speech degrades significantly when applied to NB speech, while PAD optimisation can improve performance. A higher frequency resolution might be needed for the detection of replay attacks whereas higher time resolution is need for the detection of voice conversion and speech synthesis attacks. In the face of channel variation, performance again degrades significantly. These findings show the need for new, common databases of spoofed speech which incorporate channel variation in addition to new research in channel compensation for PAD.

Acknowledgements. The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

References

[EKY13] Evans, N.; Kinnunen, T.; Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification. In: Proc. INTERSPEECH. pp. 925–929, 2013.

[GGS15] Galka, J.; Grzywacz, M.; Samborski, R.: Playback attack detection for text-dependent speaker verification over telephone channels. Speech Communication, 67:143 – 153, 2015.

[Ha15] Hautamki, R.G.; Kinnunen, T.; Hautamki, V.; Laukkanen, A.-M.: Automatic versus human speaker verification: The case of voice mimicry. Speech Communication, 72:13 – 31, 2015.

[Ha16] Hanilçi, C.; Kinnunen, T.; Sahidullah, M.; Sizov, A.: Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. Speech Communication, 85:83 – 97, 2016.

[HH15] Hansen, J. H. L.; Hasan, T.: Speaker Recognition by Machines and Humans: a tutorial review. IEEE Signal Processing Magazine, 32(6):74–99, 2015.

[IS16] ISO: ISO/IEC 30107-3: Information technology – biometric presentation attack detection. Standard, International Organization for Standardization, 2016.

[Ki17a] Kinnunen, T.; Evans, N.; Yamagishi, J.; Lee, K.-A.; Sahidullah, M.; Todisco, M.; Delgado, H.: , ASVspoof 2017: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. `http://www.asvspoof.org/data2017/asvspoof_2017_evalplan_v0.pdf`, 2017.

[Ki17b] Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.-A.: The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In: Proc. INTERSPEECH. Stockholm, Sweden, 2017.

[Ki17c] Kinnunen, T.; Sahidullah, Md.; Falcone, M.; Costantini, L.; González-Hautamäki, R.; Thomsen, D.; Sarkar, A. K.; Tan, Z.-H.; Delgado, H.; Todisco, M.; Evans, N.; Hautamäki, V.; Lee, K.-A.: RedDots Replayed: A New Replay Spoofing Attack Corpus for Text-dependent Speaker Verification Research. In: Proc. ICASSP. 2017.

[KL10] Kinnunen, T.; Li, H.: An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. Speech Communication, 52(1):12–40, Jan. 2010.

[Le15] Lee, K.-A.; Larcher, A.; Wang, G.; Kenny, P.; Brummer, N.; van Leeuwen, D.; Aronowitz, H.; Kockmann, M.; Vaquero, C.; Ma, B.; Li, H.; Stafylakis, T.; Alam, J.; Swart, A.; Perez, J.: The RedDots data collection for speaker recognition. In: Proc. INTERSPEECH. Dresden, Germany, pp. 2996–3000, 2015.

[PP15] Patel, T. B.; Patil, H. A.: Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In: Proc. INTERSPEECH. pp. 2062–2066, 2015.

[RB79] Radocy, R. E.; Boyle, J. D.: Psychological foundations of musical behavior. C. C. Thomas, 1979.

[RR95] Reynolds, D. A.; Rose, R. C.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech and Audio Processing, 3:72–83, January 1995.

[SKH15] Sahidullah, Md.; Kinnunen, T.; Hanilçi, C.: A comparison of features for synthetic speech detection. In: Proc. INTERSPEECH. Dresden, Germany, pp. 2087–2091, 2015.

[TDE16] Todisco, M.; Delgado, H.; Evans, N.: A New Feature for Automatic Speaker Verification Anti-Spoofing: constant Q Cepstral Coefficients. In: Odysey - the Speaker and Language Recognition Workshop. Bilbao, Spain, 2016.

[TDE17] Todisco, M.; Delgado, H.; Evans, N.: Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language, 2017.

[Ti16] Tian, X.; Wu, Z.; Xiao, X.; Chng, E.-S.; Li, H.: An Investigation of Spoofing Speech Detection Under Additive Noise and Reverberant Conditions. In: Proc. INTERSPEECH. pp. 1715–1719, 2016.

[VL11] Villalba, J.; Lleida, E.: Preventing replay attacks on speaker verification systems. In: IEEE International Carnahan Conference on Security Technology (ICCST). pp. 1–8, 2011.

[Wu15] Wu, Z.; Evans, N.; Kinnunen, T.; Yamagishi, J.; Alegre, F.; Li, H.: Spoofing and countermeasures for speaker verification: A survey. Speech Communication, 66:130 – 153, 2015.

[Wu17] Wu, Z.; Yamagishi, J.; Kinnunen, T.; Hanilci, C.; Sahidullah, M.; Sizov, A.; Evans, N.; Todisco, M.; Delgado, H.: ASVspoof: the Automatic Speaker Verification Spoofing and Countermeasures Challenge. IEEE Journal of Selected Topics in Signal Processing, PP(99):1–1, 2017.

[YB78] Youngberg, J.; Boll, S.: Constant-Q signal analysis and synthesis. In: Proc. ICASSP. volume 3, pp. 375–378, Apr 1978.

[Yu16] Yu, H.; Sarkar, A. K.; Thomsen, D. A. L.; Tan, Z.-H.; Ma, Z.; Guo, J.: Effect of Multi-condition Training and Speech Enhancement Methods on Spoofing Detection. In: Proc. International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE). 2016.

BIOSIG 2017

Further Conference Contributions

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 187

On the Generalization of Fused Systems in Voice Presentation Attack Detection

André R. Gonçalves[1], Pavel Korshunov[2], Ricardo P.V. Violato[1],
Flávio O. Simões[1], Sébastien Marcel[2]

Abstract: This paper describes presentation attack detection systems developed for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017). The submitted systems, using calibration and score fusion techniques, combine different sub-systems (up to 18), which are based on eight state of the art features and rely on Gaussian mixture models and feed-forward neural network classifiers. The systems achieved the top five performances in the competition. We present the proposed systems and analyze the calibration and fusion strategies employed. To assess the systems' generalization capacity, we evaluated it on an unrelated larger database recorded in Portuguese language, which is different from the English language used in the competition. These extended evaluation results show that the fusion-based system, although successful in the scope of the evaluation, lacks the ability to accurately discriminate genuine data from attacks in unknown conditions, which raises the question on how to assess the generalization ability of attack detection systems in practical application scenarios.

Keywords: Presentation attack detection, spoofed speech, cross-database evaluation.

1 Introduction

Presentation (or replay) attacks can be considered as one of the major obstacles preventing the adoption of speaker recognition in practical applications. This type of attack is relatively easy to perform. If an attacker has access to a speech sample from a target user, he/she can replay it using a loudspeaker or a smartphone to the biometric system during the authentication process. The ease of perpetration and the fact that no technical knowledge of the biometric system is required makes the presentation attack one of the most common practical attacks. Despite the severity of the problem, researchers started to develop effective presentation attack detection mechanisms only in the last few years [SKH15].

One of the main challenges in Presentation Attack Detection (PAD) is to find a set of features that allows systems to effectively distinguish speech signals that were directly emitted by a human vocal apparatus from those reproduced by a replay device such as a loudspeaker or a smartphone. Several audio descriptors originally proposed for speaker verification and speech recognition have also been studied in the context of PAD systems [SKH15] (and references there in). Features specifically designed for anti-spoofing systems were the focus of recent research [CRS07, TDE16, MMDM16].

[1] CPqD, Brazil. {andrerg,rviolato,simoes}@cpqd.com.br
[2] Idiap Research Institute, Switzerland. {pavel.korshunov,sebastien.marcel}@idiap.ch

Generalization ability of PAD systems has been assessed recently with [TDE17] showing the degradation in performance when specific features optimized using one database are used unchanged on another database. In [PSS17], cross-database experiments demonstrated the inability of current techniques to deal with unforeseen conditions. However, it did not include strict presentation attacks, which can be considered one of the hardest attack to be detected. The authors of [KM16, KM17] focused on presentation attacks in cross-database and cross-attack scenarios, and concluded that current state of the art PAD systems do not generalize well, with especially poor performance on presentation attacks.

In this paper, we present two PAD systems developed for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2017) [Ki17b]. The submitted systems are essentially ensembles of several sub-systems composed of state-of-the-art features in PAD systems and two well known classifiers: Gaussian Mixture Models (GMM) and feed-forward neural networks. Calibration and fusion strategies were used to effectively integrate these sub-systems into a possibly more robust PAD systems. We discuss and compare three different fusion strategies and investigate their performances on the ASVspoof 2017 database, as well as, by using an unrelated and larger database recorded in Portuguese language: BioCPqD-PA [Vi13] database.

2 Database and Protocol

Two different databases were used: ASVspoof 2017 and BioCPqD-PA, containing genuine and spoofed recordings. The protocol defined in the ASVspoof challenge splits the database into three subsets, while BioCPqD-PA is used as just one set. Table 1 summarizes both datasets. The databases and protocols are described in the following subsections.

Tab. 1: Number of speakers and utterances in ASVspoof 2017 and BioCPqD-PA databases.

	ASVspoof 2017			BioCPqD-PA
	train	*dev*	*eval*	-
# speakers	10	8	NA	222
# genuine	1,508	760	1,298	27,253
# spoofed	1,508	950	12,008	42,768
# total	3,016	1,710	13,306	70,021

2.1 ASVspoof 2017

The ASVspoof 2017 contest focuses on presentation attacks. To this end, the challenge makes use of the RedDots corpus [Le15] and a replayed version of the same data [Ki17a]. While the former serves as genuine samples, the latter is used as spoof samples, collected by replaying a subset of the original RedDots corpus utterances using different loudspeakers and recording devices, in different environments, through a crowdsourcing approach.

The database was split into three subsets: *train* for *training*, *dev* for *development*, and *eval* for *evaluation*. It was also allowed to use both *train* and *dev* subsets to train the final system

for score submission. The evaluation metric adopted was the Equal Error Rate (EER) and there was no need for participants to provide a decision threshold. The only restriction concerning the score was that higher scores should favor the genuine hypothesis and lower scores the spoof hypothesis. A detailed description of the contest can be found in the challenge evaluation plan [Ki17b].

2.2 BioCPqD-PA

BioCPqD-PA [Vi13] is a proprietary database that contains videos (image and audio) of participants recorded on different devices (laptops and smartphones) and environments. All recordings are in Portuguese language. The recordings (genuine audios) are from 222 speakers, collected with 4 different laptops, in 3 distinct environments, and during 5 recording sessions. In each session, 27 utterances with variable content were recorded. The total of genuine audios is 27,253.

To create the spoof attacks, a subset of these recordings (1,782 utterances sampled from all speakers' utterances in such a way that all speakers were represented in the subset) were replayed in an acoustically isolated room, using 3 different microphones and 8 different loudspeakers, resulting in 24 configurations. Then, the total number of spoofed recordings is 42,768 samples (see Table 1). In the cross-database experiments, BioCPqD-PA was used as one set. Therefore, systems tuned and trained on the ASVspoof 2017 database (following its protocol) were evaluated on the entire BioCPqD-PA, and, likewise, a system with the same configuration was trained on BioCPqD-PA and tested on ASVspoof 2017.

3 Description of the submitted PAD systems

In this section, we describe the components that constitute the two submitted PAD systems referred to as *System-1* and *System-2* in the rest of the paper.

3.1 Features

We evaluated the performance of the following features previously investigated in the context of spoofing attacks with synthetic speech: MFCC, IMFCC, RFCC, LFCC, PLP-Cepstral, SCMC, and SSFC. The use of these features was inspired by [SKH15]. Feature implementations are available online[3], which contribute to reproducibility of results. Other features were considered, such as CQCC [TDE16], PNCC [KS16], and GFCC [VA12], but in our previous tests they did not improve the performance of the jointly fused PAD system.

Features are extracted from 20ms speech frames with 50% overlap. All features are based on short-term power spectrum and were considered 20 coefficients along with their delta and delta-delta dynamic coefficients.

[3] `http://cs.joensuu.fi/~sahid/codes/AntiSpoofing_Features.zip` and in Bob framework `https://www.idiap.ch/software/bob/`

3.2 Classifiers

Two distinct classifiers were employed: a traditional 2-class Gaussian Mixture Model (GMM) classifier, where two 512 components GMM were trained (10 EM iterations), one for each class, and a Feed Forward Neural Network (FFNN), with the following architecture: Input $d \times 1 \rightarrow$ fully connected (12 neurons ReLU) $\rightarrow$ Batch Normalization $\rightarrow$ Dropout ($p = 0.5$) $\rightarrow$ fully connected (64 neurons ReLU) $\rightarrow$ Dropout ($p = 0.5$) $\rightarrow$ Sigmoid output. The cross-entropy cost function was minimized via Stochastic Gradient Descent with learning rate equals to 1e-4 with Nesterov's acceleration.

For 2-class GMM implementation, we used the system provided by the organizers with the baseline system[4] and the implementation in Bob framework[5], while FFNN was implemented in python using theano/keras[6] framework.

It is important to point out that testing different classification techniques was beyond the scope of our work for this evaluation. Therefore, a lot of space remains for assessing the use of more elaborate classifiers for PAD.

3.3 Calibration and Fusion

We focus on a score level fusion due to its relative simplicity and evidence that it leads to a better performance. The score-fusion is performed by combining scores from each of the N systems into a new feature vector of length N that needs to be classified. For classification we consider three different algorithms: (*i*) a logistic regression (LR), (*ii*) a multilayer perceptron (MLP), and (*iii*) a simple average function (Avg), which is taken on scores of the fused systems. For LR and MLP fusion, the classifier is pre-trained on the score-feature vectors from the training set.

When analyzing, comparing, and especially fusing PAD systems, it is important to have calibrated scores. Raw scores can be mapped to log-likelihood ratio scores with logistic regression, and an associated cost of calibration C_{llr} together with a discrimination loss C_{llr}^{min} are then used as application-independent performance measures of calibrated PAD or ASV systems. Calibration cost C_{llr} can be interpreted as a scalar measure that summarizes the quality of the calibrated scores. A well-calibrated system has $0 \leq C_{llr} < 1$ and produces well-calibrated likelihood ratio. Discrimination loss C_{llr}^{min} can be viewed as the theoretically best C_{llr} value of an optimally calibrated systems. We refer to [Ma14] for a discussion on the score calibration and C_{llr} and C_{llr}^{min} metrics.

3.4 Submitted systems

The two submitted PAD systems are essentially ensembles of different combinations of features and classifiers. Table 2 shows the set of sub-systems and the fusion method used

[4] `http://www.ASVspoof.org/data2017/baseline_CM.zip`
[5] `https://gitlab.idiap.ch/bob/bob.bio.gmm`
[6] Theano: `https://github.com/Theano/Theano` and Keras: `https://keras.io/`

for each PAD system. Features are presented with a subscript '*all*' or 'Δs', where '*all*' means that all static and dynamic (delta and delta-delta) features were used, while 'Δs' indicates that only the dynamic features were considered. The choice of the set of sub-systems was based on their performances measured on contest's *dev* set prior to the sub-mission.

Tab. 2: Description of the submitted systems: *System-1* and *System-2*.

	System-1	System-2
Sub-systems	**GMM** with: RFCC$_{all}$, RFCC$_{\Delta s}$, LFCC$_{all}$, LFCC$_{\Delta s}$, MFCC$_{all}$, MFCC$_{\Delta s}$, IMFCC$_{all}$, MFCC$_{\Delta s}$, SSFC$_{all}$, SSFC$_{\Delta s}$, SCMC$_{all}$, SCMC$_{\Delta s}$ **FFNN** with: IMFCC$_{all}$, LFCC$_{all}$, MFCC$_{all}$, PLP-Cepstral$_{all}$, RFCC$_{all}$, SCMC$_{all}$	**GMM** with: RFCC$_{all}$, RFCC$_{\Delta s}$, LFCC$_{all}$, LFCC$_{\Delta s}$, MFCC$_{all}$, MFCC$_{\Delta s}$, IMFCC$_{all}$, IMFCC$_{\Delta s}$, SSFC$_{all}$, SSFC$_{\Delta s}$, SCMC$_{all}$, SCMC$_{\Delta s}$
Fusion	Logistic Regression	Logistic Regression

4 Results on the ASVspoof2017 database

Table 3 shows the performance of the submitted systems in terms of EER, both for the *dev* and the *eval* sets. The results obtained for the *dev* set are based on the systems trained exclusively on the *train* set of ASVspoof2017 database, while to obtain the results for *eval* set, the systems were trained on the aggregated set: *train+dev*.

Additionally, the table shows the results of baseline system provided by the challenge organizers, which is based on CQCC front-end and 2-class GMMs back-end. *Best individual* system corresponds to a single IMFCC-based sub-system trained using GMM, which demonstrated the best performance during pre-submission evaluations. A detailed analysis of the results can be found in [Ki17b], where the results from all participants are compared.

Tab. 3: EER results for the systems submitted to ASVspoof2017, the baseline system, and the best individual model (GMM with IMFCC). The performance degradation in the Eval set is possibly due to the presence of unknown attacks. Ensemble models (System-1 and System-2) are more robust than individual models on the unseen conditions in the Eval set. Best results are highlighted.

	System-1	System-2	Best individual	Baseline
Dev (train only)	**4.09**	4.32	4.86	11.17
Eval (train+dev)	**14.31**	14.93	29.41	24.65

The only difference between baseline and best individual system is the features used, as the classifier is the same. An interesting result is the one obtained with best individual system. While on the *dev* set it provides comparable performance to the fusion-based systems, on the *eval* set it performs dramatically worse.

5 Cross-database analysis

To asses the real ability of the systems trained on the challenge database we applied them to the completely unrelated BioCPqD-PA database.

Tab. 4: EER results for the cross-database experiments: system trained on ASVspoof 2017 (*train+dev*) and tested on BioCPqD-PA, and system trained on BioCPqD-PA and tested on ASVspoof 2017 (*eval*). Best results are highlighted.

	System-1			System-2			Best
	Avg	LR	MLP	Avg	LR	MLP	individual
ASVspoof (train+dev) → BioCPqD-PA	23.35	**21.35**	22.34	22.23	**21.28**	22.41	37.24
BioCPqD-PA → ASVspoof (eval)	31.86	**26.58**	30.77	27.74	27.96	28.37	27.77

Table 4 shows that the systems trained on the ASVspoof2017 challenge database (*train+dev*) and tested on BioCPqD-PA database led to twice larger EER compared to when the same systems are evaluated on the *eval* set of ASVspoof2017 (see Table 3). This finding confirms the limited generalization power of the systems. The performance degradation in cross-database experiments is not unprecedented: it has been observed in previous anti-spoofing evaluations [TDE17, PSS17, KM16].

Three different fusion methods using Average, LR, and MLP algorithms were tested with comparable performances. LR led to a slightly better performance, especially for *System-1* trained on BioCPqD-PA database and evaluated on ASVspoof. Comparing the best individual sub-systems against fused systems, although fusion did not improve results for systems trained on BioCPqD-PA database, there is a significant improvement when it is trained on ASVspoof database. Thus, we can reason that, in practice, when the scenario is unknown, fusion add robustness to the system performance.

Observing the non-negligible difference between the two crossing possibilities in Table 4, one can arguably say that training data diversity matters. While ASVspoof database has few speakers (only male) and a limited number of utterances, it contains presumably more diverse conditions (devices and recording environments) than BioCPqD-PA, due to the crowdsourcing data collection. On the other hand, BioCPqD-PA is larger, both in terms of speakers and number of utterances, but recording conditions are more restricted.

6 Discussion

In every challenge, such as ASVspoof or NIST SRE (Speaker Recognition Evaluation[7]), the discussion about the provided speech databases emerges. Todisco et al. [TDE17] discuss the problem of selecting the features set based on results in one database and using it on another set, pointing out the resulting performance degradation. Based on our experiments, we raise another question regarding the generalization capability of systems to completely unseen conditions (including different language). Such situation is more likely to happen in practical PAD systems, where the system is trained on a given database and the attacks come from completely unknown conditions.

[7] https://www.nist.gov/itl/iad/mig/speaker-recognition

One should note that our cross-database experiments were designed for an extremely mismatched situation, when even the language is different between databases. It is expected that a PAD system should not be sensitive to language mismatch, however that might not be the case in practice, as most speech features represent acoustic properties of speech that are indeed affected by the language spoken. This has been a concern for the speaker recognition community as well: the effect of language mismatch has been evaluated in speaker recognition tasks within NIST SRE along the years.

Training a system with good generalization capability might require a larger and more diverse database. Modern algorithms based on deep learning [GBC16] approaches, for instance, which have proven to beat standard approaches in different kinds of tasks, such as speech recognition and computer vision, need massive amounts of data to provide state-of-the-art performance. In cases when the acquisition of such an amount of data is unfeasible, data augmentation strategies, such as [GUNY15], should be considered.

Another point that leads to a controversy is the use of so-called *megafusion* strategies. Although the fusion of many systems, sometimes more than a dozen (e.g., the submitted *System-1* is a fusion of 18 systems), usually leads to a better performance, its practical use is questionable. Megafusion has also been frequently used for the speaker recognition task, holding the current state-of-the-art results. However, its computational burden makes it unacceptable in practical cases, specially when system's response time is crucial.

7 Conclusions

We presented the attack detection systems developed for the Automatic Speaker Verification Spoofing and Countermeasures Challenge. The two systems achieved top five error rate (in terms of equal error rate) among 48 participants. In addition, experiments are expanded to cross-database scenario (supposedly closer to a realistic application), using BioCPqD-PA, a different unrelated database. In these experiments, a significant degradation in performance of the submitted attack detection systems is observed, highlighting the lack of generalization ability of such systems.

To improve performance, other classifiers, such as support vector machine, random forest, and deep neural networks (DNNs), need to be tested in the future. As high-generalization capability classifiers such as DNNs require a large amount of supervised training data, new data collections or data augmentation strategies will also be considered in future works. Other features specifically designed for presentation attack also need to be investigated.

Acknowledgements

This work was partially funded by Norwegian SWAN project, EU H2020 project TeSLA, and Swiss Center for Biometrics Research and Testing.

References

[CRS07] Chakroborty, J.S.; Roy, A.; Saha, G.: Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks. International Journal of Signal Processing, 4(2):114–122, 2007.

[GBC16] Goodfellow, I.; Bengio, Y.; Courville, A.: Deep Learning. MIT Press, 2016. `http://www.deeplearningbook.org`.

[GUNY15] Gonçalves, A. R.; Uliani Neto, M.; Yehia, H. C.: Accelerating replay attack detector synthesis with loudspeaker characterization. In: 7th Symposium of Instrumentation and Medical Images / 6th Symposium of Signal Processing of UNICAMP. 2015.

[Ki17a] Kinnunen, T.; Sahidullah, M.; Falcone, M.; Costantini, L.; Hautamäki, R. G.; Thomsen, D.; Sarkar, A.; Tan, Z.H.; Delgado, H.; Todisco, M.; Evans, N.; Hautamäki, V.; Lee, K.A.: RedDots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research. In: ICASSP. pp. 5395–5399, 2017.

[Ki17b] Kinnunen, T.; Sahidullah1, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A.: The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In: Interspeech. 2017.

[KM16] Korshunov, P.; Marcel, S.: Cross-database evaluation of audio-based spoofing detection systems. In: Interspeech. pp. 1705–1709, 2016.

[KM17] Korshunov, P.; Marcel, S.: Impact of Score Fusion on Voice Biometrics and Presentation Attack Detection in Cross-Database Evaluations. IEEE Journal of Selected Topics in Signal Processing, 11(4):695–705, June 2017.

[KS16] Kim, C.; Stern, R. M.: Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(7):1315–1329, 2016.

[Le15] Lee, K.; Larcher, A.; Wang, G.; Kenny, P.; Brümmer, N.; van Leeuwen, D. A.; Aronowitz, H.; Kockmann, M.; Vaquero, C.; Ma, B.; Li, H.; Stafylakis, T.; Alam, M. J.; Swart, A.; Perez, J.: The reddots data collection for speaker recognition. In: Interspeech. pp. 2996 – 2091, 2015.

[Ma14] Mandasari, M. I.; Günther, M.; Wallace, R.; Saeidi, R.; Marcel, S.; van Leeuwen, D. A.: Score calibration in face recognition. IET Biometrics, 3(4):246–256, 2014.

[MMDM16] Muckenhirn, H.; Magimai-Doss, M.; Marcel, S.: Presentation Attack Detection Using Long-Term Spectral Statistics for Trustworthy Speaker Verification. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany, pp. 1–6, Sept 2016.

[PSS17] Paul, D.; Sahidullah, M.; Saha, G.: Generalization of spoofing countermeasures: A case study with ASVspoof 2015 and BTAS 2016 corpora. In: ICASSP. pp. 2047–2051, 2017.

[SKH15] Sahidullah, M.; Kinnunen, T.; Hanilçi, C.: A Comparison of Features for Synthetic Speech Detection. In: Interspeech. pp. 2987 – 3000, 2015.

[TDE16] Todisco, M.; Delgado, H.; Evans, N.: A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients. In: Odyssey, The Speaker and Language Recognition Workshop. pp. 283–290, 2016.

[TDE17] Todisco, M.; Delgado, H.; Evans, N.: Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language, 2017.

[VA12] Valero, X.; Alias, F.: Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Transactions on Multimedia, 14(6):1684–1689, 2012.

[Vi13] Violato, R.P.V.; Neto, M. Uliani; Simões, F.O.; Pereira, T.F.; Angeloni, M.A.: BioCPqD: uma base de dados biométricos com amostras de face e voz de indivíduos brasileiros. Cadernos CPqD Tecnologia, 9(2):7–18, 2013.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 195

Exploring Texture Transfer Learning via Convolutional Neural Networks for Iris Super Resolution

Eduardo Ribeiro [1,2], Andreas Uhl [2]

Abstract: Increasingly, iris recognition towards more relaxed conditions has issued a new super-resolution field direction. In this work we evaluate the use of deep learning and transfer learning for single image super resolution applied to iris recognition. For this purpose, we explore if the nature of the images as well as if the pattern from the iris can influence the CNN transfer learning and, consequently, the results in the recognition process. The good results obtained by the texture transfer learning using a deep architecture suggest that features learned by Convolutional Neural Networks used for image super-resolution can be highly relevant to increase iris recognition rate.

Keywords: Single-Image Super Resolution, Iris Recognition, Transfer Learning, Convolutional Neural Networks.

1 Introduction

Iris recognition is one of the most accurate biometric modality for human identification mainly because of the intrinsic randomic and stable nature of the iris texture besides its high degree of freedom and noninvasive acquisition [Hs16]. In an effort to solve the problems related to the resolution of images mainly due to the iris capture distance and the inclusion of mobile devices in this field, researchers have focused on improving the image resolution that may allow the iris recognition of low resolution images since there is a substantial performance decrease directly related to the lack of pixel resolution. [Ka10]

One of the most relevant areas related to this problem is the Single-Image Super Resolution, which aim to recover a high-resolution image from a low resolution one. Examples are the use of internal patch recurrence [HSA15], regression functions [Li15] [TDV15] and sparse dictionary methods [Ya12]. However, the use of SR techniques for biometric systems especially for iris recognition is still limited including methods based on PCA eigen-patch transformation [AFFB15] and non-parametric Bayesian dictionary learning [Al15].

Over recent years, new techniques applying deep learning have been widely explored to map models from low resolution to high resolution patches primarily based in previous models applied to image denoising. Some examples are the use of Convolutional Neural Networks and Autoencoders [JAL16], [Le16], [Sh16]. Among these several successful examples, two approaches have become very popular: first the Super-Resolution Convolutional Neural Network (SRCNN) presented by [Do16] that became to be a good alternative in the first experiments for an end-to-end approach in super-resolution using Convolutional Neural Networks and then the Very Deep Convolutional Networks for Super-Resolution

This research was partially supported by CNPq-Brazil for Eduardo Ribeiro under grant No. 00736/2014-0.
[1] Federal University of Tocantins, Department of Computer Sciences, Tocantins, Brazil, uft.eduardo@uft.edu.br
[2] University of Salzburg, Department of Computer Sciences Salzburg, Austria, uhl@cosy.sbg.ac.at

(VDCNN) proposed by [KLL16] inspired by the VGG-net used for the ImageNet classification [SZ14] increasing the network depth to achieve better accuracy.

Some studies show that the use of transfer learning (approach used to improve the performance of machine learning by harnessing the knowledge acquired in another task) also can be used to assist in the task of single image super resolution as in [YZL17], [SZJ16] and [SH17]. The main problem is to know which database is more suitable to perform this transfer learning and to be able to learn the correct patterns that will be useful in the target database.

For this, in this work we aim to answer the following questions: is the similarity of the dataset used in the transfer learning important to a better mapping? Are different Iris Databases more feasible for transfer learning applied to Iris Super Resolution? In particular, do we get better results in applying the transfer learning for Super Resolution when the CNN is trained with natural image datasets, texture datasets or iris datasets? Another issue that we aim to test is if, in a practical application, we could use enrollment images in high definition already stored on the system to train a CNN and transfer the knowledge from this dataset to the entire database in order to increase accuracy of the results.

2 Methodology

2.1 Target/Test Database

To test the transfer learning with the different training databases, the chosen target database was the public iris dataset CASIAIrisV3-Interval that is the most widely use on biometrics experiments containing a total of 2.655 NIR images of size 280x320 pixels, from 249 subjects captured with a self-developed close-up camera, resulting in 396 different eyes.

In a pre-processing step, all images from this database are resized via bicubic interpolation to have the same sclera radius, then a square region of 231x231 around the pupil center is cropped. The images that do not fit in this cropping are discarded. After this procedure, 1872 images from 249 users are remained in the database. For the evaluation method, we divide this resulting database into two: one containing the first three images of each user (representing the registration images) and other containing the remaining images from each user (representing the authentication images). The registration database will be one of the used databases in the training of the CNN's and the other (authentication database) will be used for all transfer learning evaluation.

2.2 Origin/Training Databases

For the CNN training, besides the use of the registration images from the Test Database as mentioned before, we use 10 different databases including four texture databases, two natural image databases and four iris databases (from the public IRISSEG-EP [Ho14] dataset) described as follows.

- Texture Databases: The Amsterdam Library of Textures (**ALOT**) with 27500 rough texture images of size 384x256 divided into 250 classes [BG09]. The Describable Texture Dataset (**DTD**) with 5640 images of sizes range betwenn 300x300 and 640x640 categorized in 47 classes [Ci14]. The Flickr Material Database (**FMD**) containing 1000 images of size 512x384 divided into 10 categories [SRA09]. The

Textures under varying Illumination, Pose and Scale (**KTH-TIPS**) database with 10 different materials containing 81 cropped images of size 200x200 in each class [Da99].

- Natural Image Databases: The **CALTECH101** Database is a natural image dataset with a list of objects belonging to 101 categories [FFFP07]. The **COREL1000** database is a natural image database containing 1000 color photographs showing natural scenes of ten different categories [RBB08].
- Iris Databases: The IIT Delhi Iris Database (**IITD**) is an Iris Database consisting of data acquired in a real environment resulting in 2240 images of size 230x240 from a digital CMOS near-infrared camera. The CASIA-Iris-Lamp (**CASIAIL**) is an Iris database collected using a hand-held iris sensor and containing 16212 images of size 320x280 with nonlinear deformation due to variations of visible illumination. The **UBIRIS** v2 Iris database is a database containing 2250 images of size 400x300 captured on non-constrained conditions (at-a-distance, on-the-move and on the visible wavelength), attempting to simulate more realistic noise factors. The **NOTREDAME** Iris Database is a collection of close-up near-infrared Iris images containing 837 images of size 640x480 with off-angle, blur, interlacing, and occlusion factors.

2.3 CNN Architectures and Frameworks

In this work, for the comparison between different databases using transfer learning we use a classical Single-Image Super Resolution approach as base called SRCNN [KLL16]. The framework of this approach consists of three steps: patch extraction/representation, non-linear mapping and reconstruction. In this method, for the training step, patches of size 33x33 (also called High Resolution (HR) patches) are extracted from the training images and used as labels for the CNN, then those same patches are downscaled in a factor of 2 and re-upscaled to the original size using bicubic interpolation being used as inputs to the network (also called Low Resolution (LR) Patches). The SRCNN architecture is composed by three convolutional layers, where: the first layer consists of 64 filters of size 9x9x1 with stride 1 and padding 0, the second layer with 32 filters of size 1x1x64 with stride 1 and padding 0, and the last layer with 1 filter of size 5x5x32 with stride 1 and padding 0. The loss function used in this case is the Mean Squared Error (MSE) and loss minimization is done using stochastic gradient descent with the standard backpropagation method [Le01].

We also decided to use the deeper CNN VDSR [SZ14] that increases significantly the depth of the network to have a better clarification of the issues raised in this work. The framework of this approach is done by the following steps: for the training, HR patches are extracted and downscaled for the factor two, three and four (LR patches) that will serve as input of the network. In the case of this approach the labels will be the residual between the LR inputs and then HR patches. The residual-learning boost the convergence and consequently, the performance of the CNN. The VDSR architecture is composed of 20 layers and the information used for reconstruction have size of 1x41x41 (much larger than the SRCNN). The training is carried out also based on the gradient descend with backpropagation [Le01] using the MatConvNet framework [VL14].

In both frameworks, for the CNN training, a subset of 150000 patches are extracted from each database to pre-train each CNN from scratch (when the CNN weights are initialized randomly) using the pre-selected databases and use them in the target database to perform the Super-Resolution.

3 Experimental Setup

In the method evaluation, to generate the reconstructed image we use the target image database: images from CASIAIrisV3-Interval that were not used in the training for the same database (registration versus authentication images) as explained in the previous section. For each transfer learning procedure the images from the authentication database are downscaled to the desired factor : 2 (115x115), 4 (57x57), 8 (29x29) and 16 (15x15) and re-upscaled using the bicubic interpolation for factor 2, then the images pass through the deep learning CNN (SRCNN or VDCNN) to reconstruct the final super-resolved image database. Therefore, in this case, to achieve the factor 2 the image will be interpolate and pass through the trained CNN just one time. To achieve greater factors, images have to pass through the procedure $\log_2(n)$ times, where n is the desired factor.

To evaluate the performance of the transfer learning approach by quality assessment algorithms we use the the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). In these two metrics, a high metric score reflects a high quality. For the quality tests, all images from the database are used in high resolution as reference images.

Besides the quality assessment performance, we also conduct recognition experiments using the USIT - University of Salzburg Iris Toolkit v2 for Iris Recognition [Ra16] with two different iris segmentation and two feature extraction methods. In the first approach the iris is segmented and unwrapped to a normalized rectangle of 64x512 pixels using the weighted adaptive Hough and ellipsopolar transform (WAHET). Then, a complex Gabor filterbank with eight different filter size and wavelength is used to extract the iris features (CG) that will be compared using the normalized Hamming distance [Ra16]. In the second approach, the iris is segmented also using the weighted adaptive Hough and ellipsopolar transform (WAHET). Then, a classical wavelet-based feature extraction is done with a selection of spatial wavelets (QSW) that will also be compared using the normalized Hamming Distance [Ra16]. In both cases, with these procedures, using the CASIAIrisV3-Interval database with 249 users containing at least five or more images per user, we obtain 5087 genuine and 1746169 impostors scores.

We compare our method with bilinear and bicubic interpolation. We are aware that this comparison is very limited, however Super-Resolution in Iris Recognition research still is a very new field and the improvement of the comparison of transfer-learning techniques will lead to a more profound and comprehensive framework to future evaluation.

4 Results

Table 1 shows the quality assessment results for the transfer learning in different databases using the SRCNN architecture for different factors: 2, 4, 8 and 16. It can be seen that all transfer learning approaches outperform the bilinear and bicubic interpolations for all

factors including bigger factors showing the resilience of the deep-learning method when image resolution decreases.

It also can be noticed that the transfer learning using texture databases perform better in terms of similarity to the original HR database than the transfer learning using iris databases. However, the results from the Casia Interval transfer learning present good results compared to the other iris databases. The best result in this case is when the CNN is trained with the DTD database especially for higher factors and the Caltech101 database for smaller factors.

				Texture Databases				Natural Image Databases			Iris Databases			
LR Size (SCALING)		Bilinear	Bicubic	ALOT	DTD	FMD	KTH TIPS	CALTECH 101	COREL 1000	IITD	CASIAIL	UBIRIS	NOTRE DAME	CASIA INTERVAL
115X115	PSNR	0.8855	0.8957	0.9481	**0.9595**	0.9509	0.9485	0.9492	0.9491	0.9483	0.9422	0.9414	0.9495	0.9502
(1/2)	SSIM	30.77	31.07	35.17	**35.87**	35.82	35.79	35.85	35.34	35.43	35.12	34.67	35.70	35.80
57X57	PSNR	0.7949	0.8089	0.8243	**0.8259**	0.8245	0.8232	0.8250	0.8255	0.8214	0.8129	0.8131	0.8216	0.8240
(1/4)	SSIM	27.99	28.67	29.20	**29.32**	29.29	29.23	29.24	28.97	29.18	29.01	28.86	29.24	29.29
29X29	PSNR	0.6956	0.7061	0.7198	0.7228	0.7157	0.7204	**0.7251**	0.7236	0.7127	0.7064	0.7085	0.7128	0.7174
(1/8)	SSIM	24.59	25.06	25.61	25.79	25.57	25.69	**25.80**	25.50	25.44	25.15	25.12	25.44	25.54
15X15	PSNR	0.6120	0.6160	0.6510	0.6544	0.6471	0.6503	**0.6557**	0.6553	0.6439	0.6406	0.6430	0.6447	0.6494
(1/16)	SSIM	20.78	20.93	23.09	**23.23**	23.07	23.04	23.21	23.05	23.01	22.67	22.69	22.97	22.95

Table 1: Results of quality assessment algorithms for different databases training with different downscaling factors (average values on the test dataset) using the SRCNN architecture comparing to the Bilinear and Bicubic approach.

In the iris recognition verification, it can be seen from Table 2 that the results present different best results among the databases as well as presents mismatch results between the quality experimental results from table 2 and the verification results. In the case of EER the best result for the factor 2 (115X115) is when the DTD database is used (accuracy of 6.07%) in accordance with the quality assessment results (PSNR and SSIM) presenting even better results than the original database (6.657% of accuracy). Nonetheless, for the factor 4 (57x57), the best result is from the bicubic interpolation even better than all the results from the factor 2 and from the original HR database results. Among the training databases, for the recognition experiments, the more consistently beneficial for the transfer learning is the KTHTIPS database especially for the factors 4 and 8. Using the enrollment images from the same target database (Casia Interval) does not lead to good recognition performances, which means that the CNN poorly memorize the patterns from the users focusing more in general patterns, mainly because the depth of the network that does not allow a high feature discrimination.

				Texture Databases				Natural Image Database			Iris Databases			
LR Size (SCALING)		Bilinear	Bicubic	ALOT	DTD	FMD	KTH TIPS	CALTECH 101	COREL 1000	IITD	CASIAIL	UBIRIS	NOTRE DAME	CASIA INTERVAL
115X115	WAHET + CG	6.32	6.39	6.50	**6.07**	6.66	7.16	6.74	6.39	6.68	6.61	6.37	6.64	6.83
(1/2)	WAHET+QSW	**3.26**	3.58	3.58	3.32	3.81	4.28	4.02	3.53	3.89	3.92	3.42	4.02	3.84
57X57	WAHET + CG	9.36	**5.81**	7.19	6.67	6.88	6.22	6.83	6.51	7.90	7.84	8.41	7.59	6.66
(1/4)	WAHET+QSW	6.10	**2.65**	4.58	3.78	4.09	3.62	3.95	3.74	5.11	5.22	5.75	4.66	3.93
29X29	WAHET + CG	36.11	42.22	32.97	32.19	36.86	**22.41**	32.88	33.81	38.19	39.88	39.75	39.15	33.89
(1/8)	WAHET+QSW	33.60	42.34	30.62	31.13	34.89	**21.75**	32.10	33.26	36.50	38.53	37.33	37.04	30.65
15X15	WAHET + CG	31.66	32.96	33.95	33.10	33.03	33.96	33.02	34.68	32.73	**28.52**	29.62	31.50	31.57
(1/16)	WAHET+QSW	30.68	32.18	32.57	32.06	31.60	33.06	31.66	33.18	31.84	**27.60**	28.02	31.25	30.17

Table 2: Verification results (EER) for different databases training for different downscaling factors using the SRCNN architecture comparing to the Bilinear and Bicubic approach. The accuracy result for the original database with no scaling is 6.65% for WAHET + CG and and 3.81% for WAHET + QSW.

With the two better databases transfer learning from both quality assessment algorithms and recognition experiments (KTHTIPS and DTD) we decide to explore the deeper network (VDCNN) comparing the results with the CASIA INTERVAL registration images transfer learning approach also using the Very deep Super Resolution CNN (VDCNN). It can be seen in the Table 3 that this architecture leads to superior results comparing to the SRCNN in the quality measures and mainly for greater factors (8 and 16) in the recognition experiments. It also can be noticed that with deeper layers, the CNN could be able to extract more specific texture patterns from the Iris boosting the performance using Casia Interval database showing much better and consistent performances with this database.

				CASIA INTERVAL		KTHTIPS		DTD	
LR Size (SCALING)		Bilinear	Bicubic	SRCNN	VDCNN	SRCNN	VDCNN	SRCNN	VDCNN
115x115 (1/2)	PSNR	0.8855	0.8957	0.9502	0.9555	0.9485	0.9493	**0.9595**	0.9540
	SSIM	30.77	31.07	35.80	**36.90**	35.79	36.17	35.87	36.56
	WAHET + CG	6.32	6.39	6.83	6.63	7.16	6.43	**6.07**	6.32
	WAHET + QSW	**3.26**	3.58	3.84	3.78	4.28	3.63	3.32	3.53
57x57 (1/4)	PSNR	0.7949	0.8089	0.8240	0.8347	0.8232	0.8256	0.8259	**0.8348**
	SSIM	27.99	28.67	29.29	29.60	29.23	29.42	29.32	**29.65**
	WAHET + CG	9.36	**5.81**	6.66	6.51	6.22	6.83	6.67	6.69
	WAHET + QSW	6.10	**2.65**	3.93	3.26	3.62	3.41	3.78	3.41
29x29 (1/8)	PSNR	0.6956	0.7061	0.7174	0.7332	0.7204	0.7252	0.7228	**0.7374**
	SSIM	24.59	25.06	25.54	26.04	25.69	25.92	25.79	**26.21**
	WAHET + CG	36.11	42.22	33.89	**17.88**	22.41	22.14	32.19	19.07
	WAHET + QSW	33.60	42.34	30.65	**16.72**	21.75	19.20	31.13	17.07
15x15 (1/16)	PSNR	0.6120	0.6160	0.6494	0.6563	0.6503	0.6494	0.6544	**0.6633**
	SSIM	20.78	20.93	22.95	23.30	23.04	22.95	23.23	**23.57**
	WAHET + CG	31.66	32.96	**31.57**	33.87	33.96	**31.57**	33.10	33.85
	WAHET + QSW	30.68	32.18	**30.17**	32.03	33.06	**30.17**	32.06	31.76

Table 3: Quality assessment (PSNR and SSIM) and verification results (WAHET + CG and WAHET + QSW) for different databases training and different downscaling factors using the SRCNN and VDCNN architectures. The accuracy result for the original database with no scaling is 6.65% for WAHET + CG and 3.81% for WAHET + QSW.

It also can be noticed with the two different architectures comparing it to the bicubic and bilinear interpolations that, specially in the SSIM measure, the biggest drop can be observed for small down sampling factors. The CassiaInterval-VDCNN and DTD-VDCNN database present in both measures (SSIM and PSNR) superior results especially for low resolution images. On the other hand, for the recognition experiments, despite the good performance for small factors there is a significant degradation when it comes to very low resolution using these two databases. It also can be seen that despite the disparity between quality and recognition results, the databases that present the best recognition results in average are the KTHTIPS-VDCNN database and the CasiaInterval-VDCNN database specially for the factors 2, 4 and 8 that the performance is not significantly degraded. We consider that a good recognition performance is better than a quality measure in this case, so it can lead to the allowance of using small size images in systems under low storage or data transmission potential for example.

5 Conclusions

Exploring deep learning for single-image super resolution to improve the performance of iris recognition still is a new research area. In this paper we explore the use of texture transfer learning for super resolution applied to low resolution images. This approach was evaluated in a subset of Casia Iris Database representing the authentication images to also

verify if the transfer learning from the registration image subset is suitable for this application. We have shown how the features from completely different nature can be transferred in the feature domain, improving the recognition performance if applied to bigger reduction factors comparing to the classical interpolation approaches.

The experiments showed that the transfer learning was successful using all databases especially for the texture databases and using a deeper architecture in an uncontrolled scenario (when both the enrollment and the authentication images are in low resolution) despite the fact that there was not a best database to be used in all factors. In future work we intend to explore the fusion between the best databases with the enrollment images to see if the results can be even better for all cases. The direction of this research can become much more practical to many real scenarios specially in real-life applications when both the malleability of capturing devices and the recognition rate are important aspects for a successful iris recognition system.

References

[AFFB15] Alonso-Fernandez, F.; Farrugia, R. A.; Bigun, J.: Eigen-patch iris super-resolution for iris recognition improvement. In: 2015 23rd European Signal Processing Conference (EUSIPCO). Aug 2015.

[Al15] Aljadaany, R.; Luu, K.; Venugopalan, S.; Savvides, M.: IRIS super-resolution via nonparametric over-complete dictionary learning. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 3856–3860, Sept 2015.

[BG09] Burghouts, G.; Geusebroek, J.: Material-specific adaptation of color invariant features. Pattern Recognition Letters, 30(3):306 – 313, 2009.

[Ci14] Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; ; Vedaldi, A.: Describing Textures in the Wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2014.

[Da99] Dana, K.; van Ginneken, B.; Nayar, S.; Koenderink, J.: Reflectance and Texture of Real-world Surfaces. ACM Trans. Graph., 18(1):1–34, January 1999.

[Do16] Dong, C.; Loy, C. C.; He, K.; Tang, X.: Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2):295–307, Feb 2016.

[FFFP07] Fei-Fei, L.; Fergus, R.; Perona, P.: Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. Comput. Vis. Image Underst., 106(1):59–70, April 2007.

[Ho14] Hofbauer, H.; Alonso-Fernandez, F.; Wild, P.; Bigun, J.; Uhl, A.: A Ground Truth for Iris Segmentation. In: 2014 22nd International Conference on Pattern Recognition. pp. 527–532, Aug 2014.

[Hs16] Hsieh, S. H.; Li, Y. H.; Tien, C. H.; Chang, C. C.: Extending the Capture Volume of an Iris Recognition System Using Wavefront Coding and Super-Resolution. IEEE Transactions on Cybernetics, 46(12):3342–3350, Dec 2016.

[HSA15] Huang, J. B.; Singh, A.; Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5197–5206, June 2015.

[JAL16] Johnson, J.; Alahi, A.; Li, Fei-Fei: Perceptual Losses for Real-Time Style Transfer and Super-Resolution. CoRR, abs/1603.08155, 2016.

[Ka10] Kalka, N. D.; Zuo, J.; Schmid, N. A.; Cukic, B.: Estimating and Fusing Quality Factors for Iris Biometric Images. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 40(3):509–524, May 2010.

[KLL16] Kim, J.; Lee, J. K.; Lee, K. M.: Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1646–1654, June 2016.

[Le01] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.: Gradient-Based Learning Applied to Document Recognition. In: Intelligent Signal Processing. IEEE Press, 2001.

[Le16] Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Aitken, A. P.; Tejani, A.; Totz, J.; Wang, Z.; Shi, W.: Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. CoRR, abs/1609.04802, 2016.

[Li15] Li, J.; Qu, Y.; Li, C.; Xie, Y.; Wu, Y.; Fan, J.: Learning local Gaussian process regression for image super-resolution. Neurocomputing, 154, 2015.

[Ra16] Rathgeb, C.; Uhl, A.; Wild, P.; Hofbauer, H.: Design Decisions for an Iris Recognition SDK. In (Bowyer, Kevin; Burge, Mark J., eds): Handbook of Iris Recognition, Advances in Computer Vision and Pattern Recognition. Springer, second edition edition, 2016.

[RBB08] Ribeiro, E.; Barcelos, C.; Batista, M.: Image Characterization via Multilayer Neural Networks. In: 2008 20th IEEE International Conference on Tools with Artificial Intelligence. volume 1, pp. 325–332, Nov 2008.

[Sh16] Shi, W.; J.Caballero; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; Wang, Z.: Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. CoRR, abs/1609.05158, 2016.

[SH17] Sun, L.; Hays, J.: Super-resolution Using Constrained Deep Texture Synthesis. CoRR, abs/1701.07604, 2017.

[SRA09] Sharana, L.; R.Rosenholtz; Adelson, E.: Material perception: What can you see in a brief glance? Journal of Vision, 9:784, 2009.

[SZ14] Simonyan, K.; Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556, 2014.

[SZJ16] Su, M.; Zhong, S.; Jiang, J.: Transfer Learning Based on A+ for Image Super-Resolution. In (Lehner, F.; Fteimi, N., eds): Knowledge Science, Engineering and Management: 9th International Conference, KSEM 2016, Passau, Germany, October 5-7, 2016, Proceedings. Springer International Publishing, Cham, pp. 325–336, 2016.

[TDV15] Timofte, R.; DeSmet, V.; VanGool, L.: A+: Adjusted Anchored Neighborhood Regression for Fast Super-Resolution. In: 12th Asian Conference on Computer Vision. Springer International Publishing, Cham, 2015.

[VL14] Vedaldi, A.; Lenc, K.: MatConvNet - Convolutional Neural Networks for MATLAB. CoRR, abs/1412.4564, 2014.

[Ya12] Yang, J.; Wang, Z.; Lin, Z.; Cohen, S.; Huang, T.: Coupled Dictionary Training for Image Super-Resolution. IEEE Transactions on Image Processing, 21(8), Aug 2012.

[YZL17] Yuan, Y.; Zheng, X.; Lu, X.: Hyperspectral Image Superresolution by Transfer Learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(5):1963–1974, May 2017.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 203

Intrinsic Limitations of Fingerprint Orientation Estimation

Patrick Schuch[1], Simon-Daniel Schulz[2], Christoph Busch[3]

Abstract: Estimation of orientation field is a crucial issue when processing fingerprint samples. Many subsequent fingerprint processing steps depend on reliable and accurate estimations. Algorithms for such estimations are usually evaluated against ground truth data. As true ground truth is usually not available, human experts need to mark-up ground truth manually. However, the accuracy and the reliability of such mark-ups for orientation fields have not been investigated yet. Mark-ups produced by six humans allowed insights into both aspects. A Root Mean Squared Error of about 7° against true ground truth can be achieved. Reproducibility between two mark-ups of a single dactyloscopic expert is at the same precision. We concluded that the accuracy of human experts is competitive to the best algorithms evaluated at FVC-ongoing.

Keywords: fingerprint recognition, orientation field estimation, accuracy, reproducibility

1 Introduction and Motivation

The Orientation Field (OF) of a fingerprint is a characteristic feature. It represents the local orientation of the papillary ridges on the fingerprint. The OFs form typical patterns (see figure 1). They are decisive for the orientation of the characteristic points of the fingerprint ridges: the minutiae. Minutiae are the most common biometric features when recognizing fingerprints. Further processing steps may use information of the OF, e.g. image enhancement and automated minutiae extraction. Thus, fingerprint Orientation Estimation (FOE) needs to be *accurate* to allow a precise processing. This makes FOE one of the most important sub-processes in biometric feature extraction from fingerprints [Ma09].
But what does it mean to have an *accurate* FOE? An accurate FOE shall not deviate significantly from the so-called *true ground truth* (GT), i.e. the actual OF. Thus one needs to know GT for a quantitative assessment of an FOE. Unfortunately, the true GT is usually unknown as one does not know the exact OF. To circumvent this lack of true GT, human experts may mark-up GT, i.e. estimate the OF manually and record the estimation.
Whenever estimations are made, they should be questioned and analyzed for their accuracy. If in addition humans perform the estimations, reproducibility and whether the humans need expertise can be a critical issue. Despite the fact that FOE is a key aspect in biometric feature extraction, neither accuracy nor reproducibility have been assessed in literature yet. This paper addresses both aspects of FOE by humans.
As a special use case we inspect the benchmark framework FVC-ongoing. It provides the one and only relevant benchmark for quantitative assessment of algorithms for FOE. This of course makes use of a human mark-up of the GT [CMT10]. Algorithms under assessment will perform FOE on given fingerprint samples and this estimation is compared to the

[1] NTNU, NBL Norwegian Biometrics Lab, Gjøvik, NO, patrick.schuch2@ntnu.no
[2] Dermalog Identification Systems GmbH, Hamburg, DE, simon.schulz@dermalog.com
[3] NTNU, NBL Norwegian Biometrics Lab, Gjøvik, NO, christoph.busch@ntnu.no

(a) Whorl (b) Arch (c) Right Loop (d) Right Loop

Fig. 1: The presence or absence of singularities significantly shapes the orientations fields and builds typical patterns. Those singularities are *cores* (yellow crosses) and *deltas* (red crosses). The green lines emphasize the flow of the ridges around those singularities. The relative positions of the singularities can vary the shape significantly within a pattern type (compare figures 1c and 1d).

GT. GT consists of triplets (x, y, θ^{GT}) representing ground truth orientation θ^{GT} at pixel locations (x, y). Let $\theta^{E}(x, y)$ be the estimated orientation at location (x, y).
Then accuracy can be measured as the *Root Mean Squared Error* (RMSE) over all N sampling points provided in a sample:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_i^{GT} - \theta^{E}(x_i, y_i))^2} \tag{1}$$

It is worth mentioning, that the benchmark performs evaluations on two datasets: one data set contains images of good quality (GQ) and the other one contains images of bad quality (BQ). Performance is therefore measured in two scalars: $AvgErr_{GQ}$ and $AvgErr_{BQ}$ representing the average RMSE over all samples on the single datasets. This splitting takes into account the obvious fact that FOE is a harder task on BQ samples than it is on GQ samples. Published results of FVC-ongoing confirm this assumption (see figure 2). It is surprising to observe that since the FVC-ongoing benchmark was started in 2010, the $AvgErr_{BQ}$ has improved significantly over time, while $AvgErr_{GQ}$ did not. This may be an indicator for some kind lower bound for RMSE which depends on the benchmark itself. Additionally, this benchmark gives the opportunity to compare the performance of humans against the performance of algorithms tested at the benchmark.
The rest of the paper is organized as follows: Related work is described briefly in Section 2. Section 3 describes our assessment on the accuracy of FOE. The findings of this paper are summarized in section 5.

2 Related Work

Some previous work on FOE is relevant for the method proposed in this paper. One of the mark-up tools used in this work was presented by Cappelli et al. [CMM09]. Lodrova et al. have proposed averaging of minutia directions for estimations form multiple experts and define thresholds when consensus on estimations is found [Lo09]. Dactyloscopic examiners were assessed on several aspects: determination of quality [Ul14][OBB15], minutia

mark-up [Ul15][Ul16], and identification decisions [Ul11][Ul12]. Oehlmann et al compared algorithms for FOE with two further measures: average deviation (as an alternative to RMSE) and percentage of area with a deviation larger than 15° [OHG15]. This bound of 15° can be considered as a threshold between a reasonable estimation and an unacceptable deviation. They proposed to use RMSE as a measure for the accuracy for FOE against GT mark-up by a human expert. Chapman et al. provided a guide for the markup of directions of minutiae [Ch13]. Capelli et al., and Turroni et al. constituted the base for the FOE benchmark at FVC-ongoing [CMT10][Tu11]. The works of Feng et al. and of Gottschlich et al. are examples, where manually marked-up OFs were used for assessment of proposed approaches [FZJ13][GMM09]. Zhao et Jain used manual markups to separate overlapping fingerprints [ZJ12].

3 Assessment

Tools We used two different tools for mark-up. Both differ in the way the mark-up is done and how the OF is constructed from the mark-up.

Tool A is called *FingerprintAnalyzer* (see figure 3a). It was kindly provided to us by the Università di Bologna. It was the same tool which was used for marking-up the GT at FVC-ongoing. The tool allows a markup at an equidistant grid. It supports the editor by giving an initial estimation for the OF at a selected mark-up point. If the editor does not agree with this estimation, the local OF can be corrected manually. The final OF is calculated as a interpolation based on the marked-up support points. Relevant support points for interpolation are the corner of surrounding triangles of a Delaunay triangulation on the support points. The output is the OF sampled at an equidistant grid of every eighth pixel.

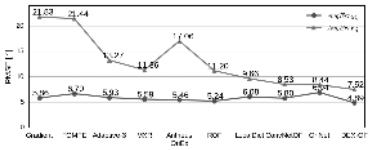


Fig. 2: Algorithms are ordered by their publication date. While $\text{AvgErr}_{\text{BQ}}$ has been improved significantly over time, $\text{AvgErr}_{\text{GQ}}$ stagnates at about 5°.

Tool B was an internal tool from our team (see figure 3b). It allows to mark-up at any point of the sample. In addition to local estimations, this tool allows to mark-up singularities (compare to figure 1). The OF is calculated as a thin plate spline (TPS) on the a complex plane based on the singularities. The global shape of the OF is modeled using a Zero-Pole Model. Local deviations from this model can be corrected using control points which use a TPS to interpolate the residual. No initial orientation proposal is provided for the control points, i.e. the orientation of the control points must be set manually. The output is an interpolated OF for every pixel.

Data Acquisition Three experts with perennial experience in the domain of fingerprints and three laymen marked-up a total of 15 samples. More reliable results would require more humans involved in the time-consuming mark-up. As we were interested in the highest achievable accuracy and best reproducibility, we focused on GQ samples. There were ten GQ fingerprint samples of dataset FOE-TEST provided by FVC-ongoing (file names are *110-119*). GT marked-up by a human was available for these ten samples.

In addition, three synthetic fingerprint samples were generated by an external synthesizer tool called SFinGe [CMM04] (see figures 3f - 3g). Two pure synthetic samples completed

(a) Tool A (b) Tool B (c) SFinGe1 (d) SFinGe2 (e) SFinGe3 (f) Lines (g) Circles

Fig. 3: Two tools are used for mark-up of OF. Tool A interpolates the OF based on marked support points (yellow lines in 3a). Additionally to local orientation (green lines), our internal tool B takes into account singularities (cyan triangles and crosses in 3b) and uses Thin Plate Splines for estimation of the OF. The set of samples to be marked-up consists of the ten GQ samples of FOE-TEST, three samples generated with SFinGE (3c-3e), and two analytical patterns (3f and 3g).

the dataset to be marked-up: straight lines and circle patterns (see figures 3f and 3g). Used frequencies were similar to those in fingerprints. For these cases true GT is available.
Both mark-up tools described in section 3 were used for mark-up (see subsection 3). Mark-up was repeated in three sessions. At least one day break was made between two consecutive sessions.
In addition, one expert performed two mark-up sessions on the 50 BQ fingerprint samples of FOE-TEST. For these samples manual marked-up GT was available, too.
Dataset FOE-TEST provided GT subsampled at an equidistant grid at every eighth pixel. This sampling rate was the lowest common denominator and was therefore used for all comparisons. In addition, the foreground area containing the fingerprint is provided with the set. As only these areas were relevant, only those were evaluated in the RMSE.

4 Analysis and Results

Accuracy The accuracy of mark-ups for FOE can be assessed most accurately only in comparison to unbiased true GT. We therefore inspected the RMSE achieved on the synthetic SFinGe samples, the lines sample and the circles sample (see figure 3). Table 1 revealed that experts performed significantly better than laymen on the task of FOE. They achieved RMSE of 7.8° for all SFinGe samples when the tool A was used. When tool B was use, 7.2° was achieved. These performances was better than the RMSE of 9.3° and 11.9° respectively achieved by the laymen. Expertise in the domain of fingerprint recognition was therefore necessary to produce a more reliable mark-up.

Person	Session	SFinGe 1	SFinGe 2	SFinGe 3	Lines	Circles	μ_{SFinGe}
Expert 1	1	8.9/7.7	8.9/8.3	6.7/5.6	1.6/0.7	6.1/2.7	**8.2/7.2**
	2	8.9/7.3	8.3/8.2	6.5/7.1	2.1/0.7	6.2/0.9	**7.9/7.5**
	3	7.7/5.3	6.9/7.9	5.2/5.7	1.7/0.7	6.7/0.7	**6.6/6.3**
Expert 2	1	8.5/6.8	8.8/7.4	8.7/6.1	1.7/0.7	8.3/1.3	**8.7/6.8**
	2	7.8/6.9	7.8/8.2	7.2/5.4	1.3/0.7	6.9/0.6	**7.6/6.9**
	3	8.5/6.1	9.0/7.8	8.6/5.9	3.4/0.7	6.6/0.6	**8.7/6.6**
Expert 3	1	9.4/9.2	10.0/7.4	6.6/6.3	2.4/0.7	5.1/0.8	**8.6/7.6**
	2	8.3/9.5	8.6/8.5	5.6/6.5	2.6/0.7	2.9/0.7	**7.5/8.2**
	3	8.3/9.6	6.6/8.7	4.7/5.9	1.5/0.7	2.2/1.1	**6.5/8.0**
Layman 1	1	12.3/21.4	9.7/13.8	7.8/19.8	2.2/0.7	4.8/6.3	**9.9/18.3**
	2	17.0/23.6	9.4/13.2	7.9/11.8	2.5/0.7	8.4/6.8	**11.5/16.2**
	3	11.7/13.0	10.8/13.5	9.5/8.0	1.9/0.7	8.1/8.7	**10.7/11.5**
Layman 2	1	10.7/11.3	7.5/12.4	7.6/8.3	1.5/2.8	5.0/7.3	**8.6/10.7**
	2	9.5/11.5	8.2/13.4	5.2/6.5	2.0/3.5	5.4/6.5	**7.6/10.5**
	3	10.8/14.7	8.2/11.2	8.3/8.4	2.5/0.0	5.4/6.6	**9.1/11.4**
Layman 3	1	10.1/9.4	9.8/10.4	5.8/8.7	1.9/2.1	7.6/5.9	**8.6/9.5**
	2	8.2/8.5	9.0/9.5	7.8/6.6	4.4/0.3	5.6/3.9	**8.3/8.2**
	3	10.6/10.8	8.5/11.1	7.9/9.1	2.6/2.8	5.6/2.6	**9.0/10.4**
$\mu_{Experts}$	**all**	**8.5/7.6**	**8.3/8.0**	**6.6/6.1**	**2.0/0.7**	**5.7/1.1**	**7.8/7.2**
$\mu_{Laymans}$	**all**	**11.2/13.8**	**9.0/12.1**	**7.5/9.7**	**2.4/1.5**	**6.2/6.1**	**9.3/11.9**
μ_{All}	**all**	**9.8/10.7**	**8.7/10.0**	**7.1/7.9**	**2.2/1.1**	**5.9/3.6**	**8.5/9.5**

Tab. 1: RMSE when marking-up with Tool A/B

The best single mark-up session for all SFinGe samples achieved RMSE of 6.2°. The RMSE achieved for the lines sample showed that this task can be performed with high

accuracy. Tool B could be used to better approximate the circles due to the capability to mark-up cores.

Gaining Expertise The development of the RMSE over the consecutive sessions gave insight, whether FOE is a task which could be learned fast. Surprisingly, laymen did not improve constantly over time. Despite this, the RMSE for the experts tended to improve over time. We assumed this effect did not reflect an improvement in the task of FOE itself. It reflected the fact that the experts got used to the tools and thus became able to express their knowledge of OF better with the tools.

Humans vs Algorithms Table 2 contains the RMSE achieved against the GT provided for the samples of FOE-TEST. This allowed to compare the performance of humans against the capabilities of those algorithms evaluated at FVC-ongoing. The mean RMSE $\mu_{110-119}$ for all experts achieved with the tool A is 6.2° and 7.0° for the tool B respectively. It is worth mentioning, that this was opposite to the higher accuracy against the true GT from the synthetic images when using tool B. This was likely due to the fact, that tool A was used to mark-up the GT. Thus, the results might slightly be biased by the mark-up tool. The best RMSE over all samples $\mu_{110-119}$ was achieved by expert 3 with the tool A: 5.2°. This was competitive to the best algorithm at FVC-ongoing (see figure 2).
As lower bounds for BQ samples were of interest, too, we performed some extra assessments. One expert additionally performed two mark-up sessions on the 50 bad quality images of dataset FOE-TEST. The expert achieved a RMSEs of 8.4° in the first and 8.3° in the second session against the alleged GT when using tool B and 11.0° and 9.6° with tool A respectively. The tool B might therefore be more appropriate for mark up of bad quality images. However, this accuracy was competitive to the best algorithm at FVC-ongoing which is called *DEX-OF* [SSBng].

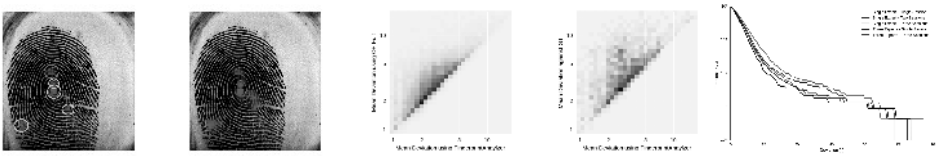
Person	Session	110	111	112	113	114	115	116	117	118	119	$\mu_{110-119}$
Expert 1	1	7.6/7.3	4.9/6.2	5.5/6.1	7.7/7.0	7.3/7.4	5.3/7.0	5.9/6.3	5.5/6.6	7.6/6.8	5.2/6.3	**6.2/6.7**
	2	6.6/5.7	6.7/6.4	6.5/5.8	7.9/7.0	8.3/6.5	5.0/6.4	7.4/5.4	5.9/7.4	7.0/7.0	5.1/5.2	**6.7/6.3**
	3	5.0/4.9	4.5/5.7	5.3/6.1	7.2/7.0	7.8/7.1	5.0/6.3	5.5/5.1	5.1/6.5	6.4/7.2	4.3/5.7	**5.6/6.2**
Expert 2	1	6.8/6.7	5.8/6.1	7.4/9.5	9.9/7.3	8.7/9.5	4.8/7.9	6.1/5.5	5.5/8.9	6.4/7.9	5.3/5.3	**6.6/7.5**
	2	6.9/5.3	4.9/6.6	5.7/9.2	8.0/6.8	7.6/8.5	4.3/6.5	5.6/6.1	5.4/7.9	6.7/8.1	5.0/5.1	**6.0/7.0**
	3	6.6/6.5	6.6/5.9	6.8/10.3	8.7/9.2	7.9/10.1	4.8/6.8	6.3/5.4	6.1/7.5	5.9/7.9	7.4/6.2	**6.7/7.6**
Expert 3	1	6.4/6.5	6.4/6.9	6.4/7.7	8.5/8.5	8.1/10.3	6.6/6.7	6.3/8.4	8.6/8.7	6.7/8.1	6.2/6.1	**7.0/7.8**
	2	5.1/7.0	4.6/7.7	4.8/6.1	6.3/8.9	6.6/7.7	4.2/5.5	5.4/6.7	7.1/8.6	5.5/8.9	4.4/5.7	**5.4/7.3**
	3	4.9/7.4	4.5/6.0	4.5/6.1	5.8/7.7	6.7/8.7	4.9/6.0	5.4/6.6	4.4/8.2	6.3/7.4	4.8/5.7	**5.2/7.0**
$\mu_{Experts}$	**all**	**6.2/6.4**	**5.4/6.4**	**5.9/7.5**	**7.8/7.7**	**7.7/8.4**	**5.0/6.6**	**6.0/6.2**	**6.0/7.8**	**6.5/7.7**	**5.3/5.7**	**6.2/7.0**

Tab. 2: RMSE against the alleged ground truth provided in dataset FOE-TEST (file names *110-119*) when marking-up with tool A/tool B. The lowest RMSE achieved over all session is 5.2°.

Local Deviations The distribution of deviations was not uniform for every sampling point. Figures 4a and 4b visualize the degree of dissent on local orientations for all experts on a single sample. Let $\theta_i^E(x,y)$ be the local estimation at location (x,y) from mark-up i. Then the local dissent $\delta(x,y)$ can be measured as the mean deviation from an averaged estimation $\mu_\theta(x,y)$ over M mark-ups:

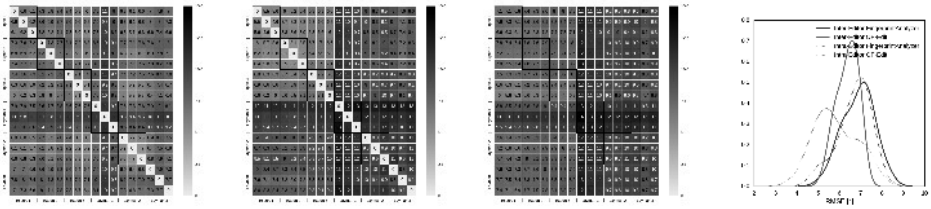
$$\mu_\theta(x,y) = 0.5 * \arctan\left(\frac{\sum_{i=1}^{M}\sin(2\cdot\theta_i^E(x,y))}{\sum_{i=1}^{M}\cos(2\cdot\theta_i^E(x,y))}\right) \tag{2}$$

$$\delta(x,y) = \frac{1}{M}\sum_{i=1}^{M}\left|\measuredangle(\theta_i^E(x,y),\mu_\theta(x,y))\right| \tag{3}$$



(a) Tool A (b) Tool B (c) Between Tools (d) Against GT (e) Averaging Mark-ups

Fig. 4: The local dissent among experts on FOE (red tinting in figures 4a and 4b) is similar for both tools. Dissent is strong near singularities (yellow circles), saddle points of curvature (blue rectangle), and where the experts need to choose between local fidelity and smoothness (green circle). Where dissent is large among the expert, the deviation to true GT is large, too (4d). Averaging over more than one mark-up can reduce such deviations (4e).



(a) Tool A (b) Tool B (c) Between Tools (d) KDE of RMSE

Fig. 5: RMSE between all sessions of all experts and laymen. The block diagonal matrix is highlighted by black squares. Those contain the comparison between all sessions of a single person and therefore allow inference on reproducibility of mark-ups.

The more intense a block was colored red, the larger was the dissent. Not surprisingly, the dissent was larger in the vicinity of singularities than it was in regions of low curvature. The local distribution of dissent was similar for both tools (see figure 4c). The area of dissent near singularities was larger for tool B than it was for tool A (yellow circles). Due to the fact that singularities could be marked-up with tool B, slight deviations in position of singularities led to larger areas of dissent. Relevant deviations can also be found where curvature has saddle points, i.e. where the ridges change their bending (blue rectangle). Additionally, there were deviations at those points, where experts had to decide between smoothness of the OF and high fidelity to local changes of the OF (green circle). This was more an individual bias than it was a critical deviation.
The local deviation among the experts from their estimated mean was strongly correlated to their mean deviation against the GT on the three samples generated with SFinGE. Pearson's correlation coefficient between both mean deviations is 0.8. Therefore, it is likely that dissent among multiple mark-ups will coincide with deviations from true GT.

Reproducibility Whenever humans are involved in processes, reproducibility is an important issue . Single mark-up sessions of the human editors were compared against each other to assess this aspect. Figures 5a-5c visualizes the RMSE between all mark-ups made by the six human editors. Since also RMSE between all sessions of a single person were

included in this graphic, it contains information regarding reproducibility. In general, experts achieved lower RMSE between their sessions than the laymen did. This holds except for layman 2 when marking up with tool A. This good reproducibility needed to be put into perspective of significant higher deviation against true GT (see table 1).
However experts could achieve RMSE between 5° and 7° between two mark-ups. Surprisingly, these accuracies were only slightly better than the accuracies between the particular experts. This was an indicator that the single mark-ups were good estimations of the true OF. The RMSE between the two sessions on the BQ samples was 11.7° when using the tool A and 7.6° when using the tool B.

Approximating True GT It seemed, that the mark-ups could be interpreted as true GT disturbed by some *noise*. If the noise is mean-free, averaging mark-ups will reduce the influence of noise. Figure 4e visualized the empirical cumulative density function of deviations between μ_θ and the true GT of the SFinGe samples. The more mark-ups involved in averaging, the lower was the deviation against the true GT. There was no significant difference between averaging all three mark-ups of one expert and averaging one session each from all three experts.

5 Conclusions

By extensive and time consuming mark-up of OFs, we investigated questions regarding FOE when performed by humans. We found that expertise in fingerprints increases the accuracy of marked-up OFs. Experts achieved an RMSE of about 7° compared to true GT. Averaging over more than one mark-up increased the accuracy. Inspection of multiple mark-ups of a single expert showed, that mark-ups could be produced at similar values of RMSE. These values were, therefore, interpreted as rough lower bounds for a reasonable accuracy at FVC-ongoing. When humans were compared to the alleged GT at benchmark FVC-ongoing, they achieved roughly 5° on GQ samples and about 8.4° on BG samples respectively. This was competitive to the best algorithms evaluated by FVC-ongoing.

Acknowledgment

The authors would like to thank the team from Università di Bologna for providing us with their mark-up tool. Additionally, we would like to thank all experts and laymen involved in the mark-up of the OFs.

References

[Ch13] Chapman, Will; Hicklin, RA; Kiebuzinski, GI; Komarinski, Peter; Mayer-Splain, John; Taylor, Melissa; Wallner, Rachel: Markup Instructions for Extended Friction Ridge Features. NIST Special Publication, 1151, 2013.

[CMM04] Cappelli, Raffaele; Maio, D; Maltoni, D: SFinGe: an approach to synthetic fingerprint generation. In: International Workshop on Biometric Technologies (BT2004). pp. 147–154, 2004.

[CMM09] Cappelli, Raffaele; Maio, Dario; Maltoni, Davide: Semi-automatic enhancement of very low quality fingerprints. In: Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on. IEEE, pp. 678–683, 2009.

[CMT10] Cappelli, Raffaele; Maltoni, Davide; Turroni, Francesco: Benchmarking local orientation extraction in fingerprint recognition. In: Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, pp. 1144–1147, 2010.

[FZJ13] Feng, Jianjiang; Zhou, Jie; Jain, Anil K: Orientation field estimation for latent fingerprint enhancement. IEEE transactions on pattern analysis and machine intelligence, 35(4):925–940, 2013.

[GMM09] Gottschlich, Carsten; Mihailescu, Preda; Munk, Axel: Robust orientation field estimation and extrapolation using semilocal line sensors. IEEE Transactions on Information Forensics and Security, 4(4):802–811, 2009.

[Lo09] Lodrova, Dana; Busch, Christoph; Tabassi, Elham; Krodel, Wolfgang; Drahansky, Martin et al.: Semantic conformance testing methodology for finger minutiae data. Proceedings of the Special Interest Group on Biometrics and Electronic Signatures (BIOSIG), pp. 31–42, 2009.

[Ma09] Maltoni, Davide; Maio, Dario; Jain, Anil K; Prabhakar, Salil: Handbook of fingerprint recognition. springer, 2009.

[OBB15] Olsen, Martin Aastrup; Bockeler, Martin; Busch, Christoph: Predicting dactyloscopic examiner fingerprint image quality assessments. In: Biometrics Special Interest Group (BIOSIG), 2015 International Conference of the. IEEE, pp. 1–12, 2015.

[OHG15] Oehlmann, Lars; Huckemann, Stephan; Gottschlich, Carsten: Performance evaluation of fingerprint orientation field reconstruction methods. In: Biometrics and Forensics (IWBF), 2015 International Workshop on. IEEE, pp. 1–6, 2015.

[SSBng] Schuch, Patrick; Schulz, Simon-Daniel; Busch, Christoph: Deep Expectation for Estimation of Fingerprint Orientation Fields. forthcoming.

[Tu11] Turroni, Francesco; Maltoni, Davide; Cappelli, Raffaele; Maio, Dario: Improving fingerprint orientation extraction. Information Forensics and Security, IEEE Transactions on, 6(3):1002–1013, 2011.

[Ul11] Ulery, Bradford T; Hicklin, R Austin; Buscaglia, JoAnn; Roberts, Maria Antonia: Accuracy and reliability of forensic latent fingerprint decisions. Proceedings of the National Academy of Sciences, 108(19):7733–7738, 2011.

[Ul12] Ulery, Bradford T; Hicklin, R Austin; Buscaglia, JoAnn; Roberts, Maria Antonia: Repeatability and reproducibility of decisions by latent fingerprint examiners. PloS one, 7(3):e32800, 2012.

[Ul14] Ulery, Bradford T; Hicklin, R Austin; Roberts, Maria Antonia; Buscaglia, JoAnn: Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PloS one, 9(11):e110179, 2014.

[Ul15] Ulery, Bradford T; Hicklin, R Austin; Roberts, Maria Antonia; Buscaglia, JoAnn: Changes in latent fingerprint examinersâ markup between analysis and comparison. Forensic science international, 247:54–61, 2015.

[Ul16] Ulery, Bradford T; Hicklin, R Austin; Roberts, Maria Antonia; Buscaglia, JoAnn: Interexaminer variation of minutia markup on latent fingerprints. Forensic science international, 264:89–99, 2016.

[ZJ12] Zhao, Qijun; Jain, Anil K: Model based separation of overlapping latent fingerprints. IEEE Transactions on Information Forensics and Security, 7(3):904–918, 2012.

Unobtrusive Gait Recognition using Smartwatches

Neamah Al-Naffakh[1,2], Nathan Clarke[1,3], Fudong Li[1], Paul Haskell-Dowland[1,3]

Abstract: Gait recognition is a technique that identifies or verifies people based upon their walking patterns. Smartwatches, which contain an accelerometer and gyroscope have recently been used to implement gait-based biometrics. However, this prior work relied upon data from single sessions for both training and testing, which is not realistic and can lead to overly optimistic performance results. This paper aims to remedy some of these problems by training and evaluating a smartwatch-based biometric system on data obtained from different days. Also, it proposes an advanced feature selection approach to identify optimal features for each user. Two experiments are presented under three different scenarios: Same-Day, Mixed-Day, and Cross-Day. Competitive results were achieved (best EERs of 0.13% and 3.12% by using the Same day data for accelerometer and gyroscope respectively and 0.69% and 7.97% for the same sensors under the Cross-Day evaluation. The results show that the technology is sufficiently capable and the signals captured sufficiently discriminative to be useful in performing gait recognition.

Keywords: mobile authentication, gait biomtrics; accelerometer; smartwatch authentication

1 Introduction

Billions of mobile devices are being used globally having a wide variety of applications (e.g., e-commerce and banking). The use of mobile devices has inherently raised security concerns and there exists a prevalent requirement to secure these devices. Smartwatches have been steadily increasing in popularity and this trend is expected to continue as the technology improves. Therefore, wearables could be used to enhance mobile security in a more effective way. Recent studies have demonstrated that both smartphones [DNBB10, MM14, NWB12] and smartwatches [JW15, SMS16, ACDL16] can provide gait-based biometric authentication service by using various sensors. However, the majority of prior research either used a limited dataset or trained and tested the system on data that was collected on the same day (which is not a realistic model for a real world application as the user would be required to enroll on the system every day). To this end, this paper explores the use of smartwatches for transparent authentication based upon gait recogntion. The main contributions of this study are demonstrated as follows

- To the best of the author's knowledge, this is the biggest dataset for smartwatch-based gait authentication, which contains gait data of 60 users over multiple days.
- A comprehensive feature set was extracted in the time and frequency domains and analyzed to highlight their impact on system performance.
- The novel feature selection method utilised a dynamic feature vector for each user and successfully reduced the feature vector size with better performance.
- Identifying the optimal source sensor for the authentication task.
- The results of this study outperform the prior biometric accelerometer –based studies.

[1] School of Computing, Electronics and Mathematics, Plymouth University, UK,
[2] Computer Science and Mathematics College, Kufa University, Najaf, Iraq
[3]Security Research Institute, Edith Cowan University, Perth, Western Australia
{Neamah.Al-Naffakh, N.Clarke, Fudong.Li}@plymouth.ac.uk; p.haskelldowland@ecu.edu.au

The rest of the paper is organized as follows: Section 2 reviews the state of the art in transparent and continuous authentication that specifically uses accelerometer (Acc) and gyroscope (Gyr) sensors. Data collection and feature extraction are outlined in Section 3. Sections 4, 5 and 6 present the experiment design, feature selection approach, results and discussion. Section 7 presents the conclusions and future research directions.

2 Related Work

Gait-based biometric systems have an advantage over password-based systems in that impersonation is much more difficult to accomplish even video footage of someone walking on a treadmill (to match the victim's pace) is not sufficient to mimic a user [GSB07]. Verifying people based on their walking patterns is an unobtrusive mechanism that does not require explicit user interaction and provide continuous authentication. Recently, increased interests are shown in mobile gait authentication; and performance rates vary considerably depending upon feature extraction methods and types of classifiers utilised. A comprehensive analysis of the prior studies on gait authentication using smartwatch and mobile sensors is summarized in Table 1.

Tab 1. Comprehensive Analysis on Gait Authentication using Mobile and Smartwatch Sensors

Study	Approach	Features Type	Classification methods	Accuracy %	Users	Duration (Seconds)	Device
[DNBB10]	C	TD	DTW	20.1 (EER)	51	120/MD	M
[NDBB11]	C	TD	DTW	21.7 (EER)	48	1200/CD	M
[MM14]	C	TD	DTW	19 (EER)	35	240/CD	M
[KWM10]	S	TD	NN	100 (CCR)	10	300-600/SD	M
[NB11]	S	FD	HMM	6.15 (EER)	48	1200/CD	M
[HN12]	S	FD	SVM	10 (EER)	36	1200/CD	M
[NWB12]	S	FD	KNN	8.24 (EER)	36	1200/CD	M
[CLB15]	S	TD	BN	96.27(CCR)	44	400/SD	M
[NBB11]	S	TD	SVM	10 (EER)	51	120/CD	M
[JW15]	S	TD	RF	1.4 (EER)	59	300-600/SD	SW
[WTGYS16]	S	TD	RF	94.2 (CCR)	17	2160/SD	SW
[CAM16]	S	TD	KNN	2.9 (EER)	15	- / SD	SW
[SMS16]	S	TD	RF	2.6 (EER)	18	350/CD	SW
[KPR16]	S	TD & FD	KNN	95 (CCR)	40	240/SD	SW
[ZYCWS17]	S	TD & FD	NN	0.5 (EER)	9	- / SD	SW

Legend: **C**: Cycle-based; **S**: Segment-based; **TD**: Time Domain; **FD**: Freqency Domain; **DTW**: Dynamic Time Warping; **HMM**: Hidden Markov Model; **SVM**: Support Vector Machine; **KNN**: k-nearest neighbours; **RF**: Random Forest; **NN** Neural Network; **EER**: Equal Error Rate; **CCR**: Correct Classification Rate; **M**: Mobile; **SW**: Smartwatch; **SD**: Same Day; **, CD**: Cross Day

Two main approaches can be used to extract gait features, namely cycle and segment-based. Cycle extraction attempts to segment the data into pairs of steps. This offers a very exciting opportunity where if such a system is implemented effectively. Howeover, the literature shows high EERs (ranging from 19% [MM14] to 21.7% [NDBB11]). This is most likely the result of the complicated and unclear nature of cycle extraction. In

contrast, the performance of the segment based methods, which focus on fixed-length blocks of data, appearing to be more effective and stable, with studies reporting EERs between 1.4% and 10 % [JW15, NBB11]. With respect to features, several studies in the domain have used both time domain (TD) and frequency domain (FD) features but little attention has given to measure the impact of these features on the system performance.
As illustrated in Table 1, the most recent studies used a smartwatch device to collect the Acc and Gyr gait data for transparent authentication systems (TAS). However, in [JW15, WTGYS16, CAM16, ZYCWS17] the gait data was obtained on the SD and the dataset is considered limited ranging from 9 to 18 users (apart from JW15). In addition, the authors did not carry out any particular study on feature selection in order to identify the most discriminative features. In contrast, a feature selection mechanism was conducted by [KPR16] and reported 95% CCR by using the SD scenario. However, the system performance was reduced to 86.8% CCR (with a limited dataset of 13 users only) when the CD scenario was applied. This can be attributed that the proposed approach is not sophisticated enough to identify a unique feature set for individuals that work over time.

3 Data Collection and Feature Extraction

The Acc and Gyr data was captured from the Microsoft Band 2 at a rate of 32 samples per second for the x, y and z-axes and automatically sent to a smartphone residing in the user's pocket via Bluetooth. In total, 60 users participated in the data collection; each user was required to walk on a predefined route in two sessions on two different days (within a time frame of 3 weeks between the sessions). Every session consisted of three walks trails from each user. In each trail, the user was asked to walk at a natural speed on flat ground for 2 minutes with few turns. For a more realistic scenario, the subject had to stop in order to open a door. Moreover, no other variables, such as type of footwear or clothing, are controlled. Once the data collection was completed, the signal processing phase was undertaken- a brief description of the steps is as follows

- Time interpolation: as the Microsoft Band 2 sensors were not able to record data at a fixed sample rate, time interpolation was required to make sure that the time period between two successive data points was always equal
- Filtering: a low pass filter was designed in order to enhance the accuracy of the signal. This was carried out with several settings (i.e. 0.1, 0.2, 0.3, 0.4 and 0.5) and the cut-off frequency of 0.2Hz achieved the best accuracy.
- Segmentation: the tri-axial raw format for both Acc and Gyr signals were segmented into 10-second segments by using a sliding window approach with no overlapping. Therefore, in total 36 samples were collected for each user per day.

A feature extraction process is carried out on both the Acc and Gyr data segments of each user. In total, 140 features were extracted based upon prior work identified in gait recognition studies. Features were extracted from both the time and frequency domains on Acc or Gyr data. Since most features are generated on a per-axis basis and each sensor has 3 axes, most features are represented by a multiple of three values. The number of generated features and their types are presented in Table 2. Details of these features (e.g., how they are calculated) can be found in [KWM10, JW15].

Tab 2. List of the extracted TD and FD features

Feature Type	NF	TD	FD	Feature Type	NF	TD	FD
Difference	3	√	√	Skewness	3	√	√
Variance	3	√	√	Average	3	√	√
Median	3	√	√	Kurtosis	3	√	√
Maximum	3	√	√	Minimum	3	√	√
Energy	3	-	√	Entropy	3	-	√
Time Between Peaks	3	√	-	Standard Deviation	3	√	√
Correlation Coefficients	3	√	√	Root Mean square	3	√	√
Cosine Similarity	3	√	-	Covariance	3	√	-
Interquartile range	3	√	√	Binned histogram	30	√	-
Peaks Occurrence	3	√	-	Percentile 25,50	6	√	√
Average Absolute Difference	3	√	√	Average Resultant Acceleration	1	√	√

4 Experimental Methodology

Biometric authentication or verification is a binary classification problem, where the aim is to determine if a system can identify a genuine user correctly or as an imposter. A separate model is generated for each user. The reference and testing templates were created under three different scenarios for SD, MD and CD. For SD and MD, the data was divided into two sets: 60% of the data for training and the remaining 40% for testing; also training samples were extracted from both days for the MD scenario. For the CD scenario, the first day's data was used for training and the second day data was employed for testing. Also, the Feedforward Multi-layer Perceptron (FF MLP) neural network was used as the default classifier due to its reliable performance [KWM10].

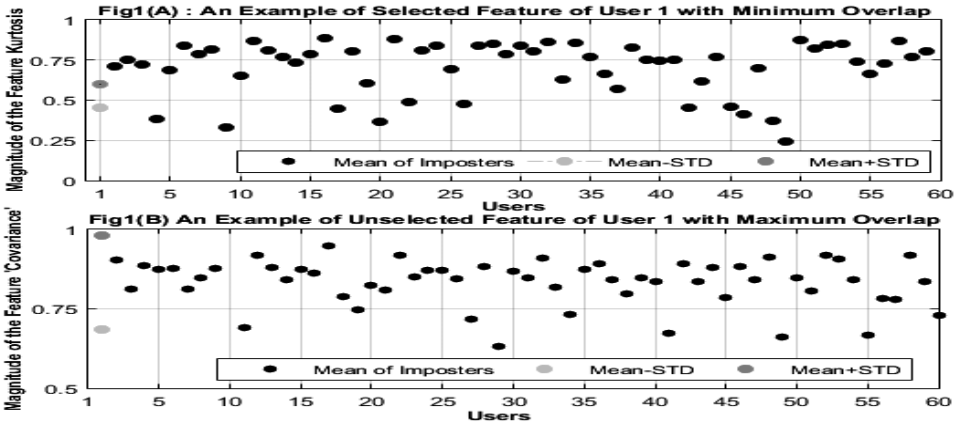
The feature selection step is important for biometrics based studies in order to reduce the potentially large dimensionality of input data. By selecting an optimal feature set for individuals, the system performance could be potentially enhanced. Also, it will be easier to manipulate and calculate smaller feature subsets on digital devices. Majority of gait recognition systems select common features for all the population; this could be useful if the system is based on identifying the genuine user only. However, a balance between security and usability needs to be taken. Therefore, this study focused on creating a dynamic feature vector that contains distinctive features for each user. As a result, the feature subset for each user very different from each other (e.g., the reference templates could be created by using features 1, 2, and 7 for user 1 while features 3, 4, and 5 for user 2). This can be achieved by calculating the mean and Standard Deviation (STD) for each feature of all users and then compares the authorized user's results against impostors to select the feature set with the minimal overlap. In other words, for each feature, a score is calculated based upon the following condition:

- If the mean of imposter's activity is not within the range of the mean +/- STD of genuine, add 1 to the total score.
- Dynamically select the features according to their score order from high to low. The highest score means less overlap between imposters and genuine user (see Fig 1 (A)).

Fig 1 shows an example of applying the proposed feature selection method on two different features for user 1. Based upon the overlap percentage, it is clear from Fig 1 that the Kurtosis feature has lowest overlap score compared to the Covariance feature.

As a result, the Kurtosis feature was selected to form the feature vector of user1, while the second feature (i.e. Covariance) was neglected. This procedure is repeated for each individual and each feature resulting in a bespoke and prioritized feature set.

Fig 1: the effect of the dynamic feature selection approach



In order to evaluate the proposed system, several consecutive experiments were undertaken that include:

- Analysis and highlighting the impact of the time and frequency domain features on the system performance.
- The discriminative features were evaluated and the reference and test templates were created by selecting an optimal feature set for each user independently.
- The results cover the three evaluation scenarios (SD, MD, and CD), the two different sensors (Acc and Gyr), and one classification algorithm (FF MLP neural network).

5 Results

According to the plan, the first expermient was to highlight the impact of the time and frequency domains features on the system performance and the results are presented in Table 3(using the SD scenario).

Table 3: EER of Using All Features, Time and Frequency Domains

Feature type	NF	EER (%)	
		Acc	Gyr
All Features	140	**0.13**	3.37
Time domain	88	**0.15**	3.73
Frequency domain	52	3.09	12.69

It is clear that good performances were achieved by using the TD features and all feature sets; and little difference in results is observed between the two sets. By using the FD features alone, reasonable performance is obtained; but its performance is far less promising in comparison with the results of using TD features alone, suggesting FD features add little contribution towards the classification process. Given the fact that detecting redundancies features makes the system more efficient, therefore, only the TD features (i.e. 88 features) were used in subsequent experiments as it shows low EER.

Further analysis was conducted to reduce the extracted TD features by applying the proposed dynamic feature selection method. Table 4 shows the impact of feature selection under the SD, MD and CD scenarios and two sensors. It can be concluded that the feature selection mechanism has a positive effect on the performance by minimizing the number of features and maximizing the discriminative information. In addition, as expected the system performance of the SD and MD scenarios exceeded the CD evaluation for both sensors.

Tab 4: Impact of the dynamic feature selection technique upon the performance in detail.

Evaluation Scenario	Sensor	Number of Selected Features								
		10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**88**
SD	Acc	1.13	0.78	0.24	0.26	0.27	**0.13**	0.20	0.16	0.15
SD	Gyr	6.6	4.88	3.63	3.74	**3.12**	3.58	3.48	3.43	3.73
MD	Acc	2.22	0.82	0.42	0.22	0.25	0.20	0.22	**0.16**	0.28
MD	Gyr	7.63	4.81	3.85	3.80	3.53	3.51	**3.24**	3.25	3.35
CD	Acc	4.68	2.39	1.43	0.9	0.84	0.83	**0.69**	0.77	0.93
CD	Gyr	11.09	9.76	8.62	8.49	8.94	8.53	8.42	**7.97**	8.29

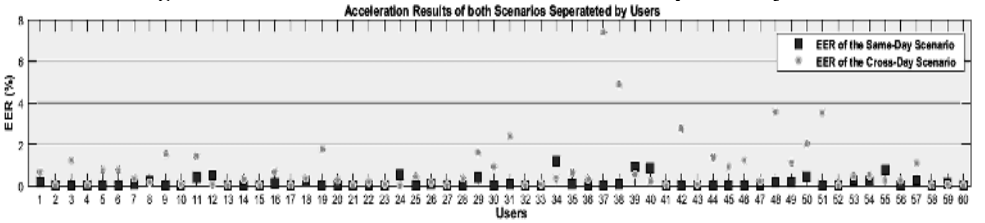
As shown in the Table 4 vastly good results were achieved with best EERs of 0.13% for Acc and 3.12% for Gyr by utilizing the SD scenario (compared to 2.9%, 1.4% and 0.5% of EERs by [CAM16, JW15, ZYCWS17] and CCR of 95% and 94% by [WTGYS16, KPR16]. Moreover, high performances with EERs of 0.78% and 4.88% can still be achieved by using only 20 features for Acc and Gyr accordingly. Comparing to the SD scenario, no significant difference was found in the MD scenario where the best EERs are 0.16% for Acc and 3.24% for Gyr, as the training set contained samples from both days. However, these results outperform the outputs (i.e. EER ranging from 6.1% to 21.7%) of previous studies [NB11, NBB11] under the MD scenario.

As shown in Table 4, the best performance of the CD scenario are EERs of 0.69% (for Acc) and 7.97% (for Gyr). As expected the system performance is droped under the CD test as the human's behaviour does change over time. Nonetheless, the presented CD results are still very promising (i.e. 0.69% EER) in comparison with the prior work that reported EERs in the range of 2.6% - 21.7% [NDBB1, MM14, NB11, HN12, NWB12, NBB11, SMS16] . In addition, the CD test does not require the user to re-enrol in the system on a daily basis.

With the aim to understand how individual user performed, results on each user's Acc for both SD and CD scenarios are presented in Figure 2. As shown in Figure 2 high level of performance (i.e. in the range of 0-2% EER) were obtained for 90% of users, (apart from users 31, 37, 38, 42, 48, and 51) for both SD and CD scenarios. This suggests that users have a consistent and distinctive set of Acc pattern characteristics.

With respect to the feature subset size, as shown in Table 4 the SD test requires less features (i.e., 60 features) than the CD (i.e., 80 features) to produce the lowest EER. This could be explained because the user's gait pattern could vary or be inconsistent over time due to many factors (e.g., shoes, clothes, and mood), hence more features are required for individual to be identified. Moreover, creating a dynamic feature vector size for each user independently might greatly reduce the EER (e.g., the refrence template can be constructed by using 20 features for user 1 while 40 features will be used for user 2)

Fig 2: The Acceleration Results of Both Scenarios Separated by Users



6 Discussion

As shown in the previous section, the presented results reveal that smartwatch based gait recognition is highly efficient and recommended to be used for verifying users in a transparent and continuous manner. The best results were EERs of 0.13% and 0.69% for SD and CD scenarios respectively by using Acc signals. However, the results were obtained in controlled conditions, so, further investigation is required by collecting the user's data during the entire day over multiple days in order to find the influence of collecting real life data on the system performance. Although features were extracted from both time and frequency domains, the findings in Table 2 support the use of time domain features alone as a better decision especially for mobile devices. For the realistic test, the EER was slightly increased from 0.13% to 0.69% when the Acc reference and test templates were created from the data of two different days. Because the obtained Acc results were very strong, the fusion of data from both sensors was not necessary. Further influencing factors on the biometric system performance is the selected feature subset; selecting unique features for each user would improve the results and reduce the complex computations on the smart devices which have limited processing resources. Therefore, a feature selection approach of any mobile-based biometric system needs to be sophisticated enough before the classification phase takes place. As expected, the proposed feature selection approach in this study, which was based on creating a dynamic feature vector for each user, successfully reduced the user's feature vector size and resulted in lower EER's of 0.13% and 0.69% for the SD and CD tests respectively (compared to 0.15% and 0.93% when the whole features were used). However, further investigation is required to reduce the number of the optimal features for each user independently which might offer better accuracy/error rates.

7 Conclusions and future work

Based on the performance in this study, smartwatch-based gait recognition shown to be effective and can be used with in TAS. The paper also presents an analysis of the feature set to examine the impact of features upon performance, which has resulted in proposing a dynamic feature set. The proposed system was evaluated by collecting the motion data from 60 users and analysed the feature set to determine its uniqueness. However, more experimental work should be carried out to investigate the impact of the dynamic feature vector size for each user.

Further work will also explore examining a wider range of different activities (e.g., fast

walking and typing on smartphone touch screen) to expand the technique from merely gait recognition to activity recognition. A future study will aim to remove the one factor that is explicitly controlled in all previous studies – the nature of the controlled data collection and instead look to understand what the performance of the approach is with real life data over a prolonged period of time. As the nature of the real life signals is likely to be noisy, an appraoch will be used in order to predict the user's activity.

References

[ACDL16] N. AI-Naffakh, N. Clarke, P. Dowland and F. Li, "Activity Recognition using Wearable Computing", in ICITST, Barcelona, pp. 189-195, 2016.

[CAM16] Cola, G., Avvenuti, M. and Musso, F.. Gait-based authentication using a wrist-worn device. in MOBIQUITOUS, Hiroshima, 2016.

[CLB15] N. Capela, E. Lemaire and N. Baddour, "Feature Selection for Wearable Smartphone-Based Human Activity Recognition with Able bodied, Elderly, and Stroke Patients", PLOS ONE, vol. 10, no. 4, p. e0124414, 2015.

[DNBB10] M. O. Derawi, C. Nickel, P. Bours and C. Busch, "Unobtrusive User-Authentication on Mobile Phones using Biometric Gait Recognition",in IIH-MSP, pp.306-311, 2010.

[GSB07] D. Gafurov, E. Snekkenes and P. Bours, "Spoof Attacks on Gait Authentication System",IEEE Transactions on Information Forensics and Security,pp. 491-502, 2007.

[HN12] M. Reese Hestbek and C. Nickel, "Biometric gait recognition for mobile devices using wavelet transform and support vector machines", in IWSSIP, pp. 205-210, 2012.

[JW15] A. H. Johnston and G. M. Weiss, "Smartwatch-Based Biometric Gait Recognition", in Biometrics Theory, Applications and Systems (BTAS), Arlington, VA, USA, 2015.

[KPR16] R. Kumar, V.V Phoha, and R. Raina: Authenticating users through their arm movement patterns. arXiv preprint arXiv:1603.02211 (2016).

[KWM10] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell phone-based biometric identification", in *the Fourth IEEE International BTAS Conference*, 2010, pp. 1–7.

[MM14] M. Muaaz and R. Mayrhofer, "Orientation Independent Cell Phone Based Gait Authentication", in MOMM Conference, Taiwan, 2014.

[NB11] C. Nickel and C. Busch,"Classifying accelerometer data via Hidden Markov Models to authenticate people by the way they walk", in ICCST, Barcelona, pp. 1–6, 2011.

[NBB11] C. Nickel, H. Brandt and C. Busch, "Classification of Acceleration Data for Biometric Gait Recognition on Mobile Devices", in BIOSIG Conference, Darmstadt, 2011.

[NDBB11] C. Nickel, M. O. Derawi, P. Bours, and C. Busch, "Scenario test of accelerometer-based biometric gait recognition", in *the Third IWSCN* Conference, pp. 15–21, 2011.

[NWB12] C. Nickel, T. Wirtl and C. Busch, "Authentication of Smartphone Users Based on the Way They Walk Using k-NN Algorithm", in IIH-MSP, Greece, pp. 16-20, 2012.

[SMS16] Babins, S., Manar, M., Nitesh, S.: Walk-Unlock: Zero-Interaction Authentication Protected with Multi-Modal Gait Biometrics. arXiv:1605.00766 (2016).

[WTGYS16] G. M. Weiss, J. L. Timko, C. M Gallagher, K. Yoneda and A. J. Schreiber, "Smartwatch-based activity recognition: A machine learning approach", in BHI, Las Vegas, USA, pp. 426-429, 2016.

[ZYCWS17] ZHANG, X., YAO, L., CHEN, K., WANG, X., Z. SHENG, Q. and GU, T: DeepKey: An EEG and Gait Based Dual-Authentication System. arXiv: 1706.01606 (2017).

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 219

Evaluation of Motion-based Touch-typing Biometrics in Online Financial Environments

Attaullah Buriro[1], Sandeep Gupta[1], Bruno Crispo[1,2]

Abstract: This paper presents a bimodal scheme, the mechanism which contemplates the way a user enters an 8-digit PIN/password and the phone-movements while doing so, for user authentication in mobile banking/financial applications (apps). The scheme authenticates the user based on the timing differences of the entered strokes. Additionally, it enhances the security by introducing a transparent layer utilizing the phone-movements made by the user. The scheme is assumed to be highly secure as mimicking the invisible touch-timings and the phone-movements could be extremely onerous. Our analysis is based on 2850 samples collected from 95 users through a 3-day unsupervised field experiment and using 3 multi-class classifiers. Random Forest (RF) classifier out-performed other two classifiers and provided a True Acceptance Rate (TAR) of 96%.

Keywords: Smartphones, Biometric Authentication, Human-Computer Interaction

1 Introduction

Mobile banking is among the most sensitive activity a user performs on the Internet. Almost every bank now offers mobile banking through their dedicated apps. Thus, increasing number of smartphone users carry their banks around in their pocket rather than limiting themselves to just desktops or laptops. Recent research revealed that more than 82% of the teenage users (between 25 to 35 years) and 70% of the household users use online banking from their smartphones[3].

Mobile banking apps perform remote authentication requiring user credentials, as proof of identities, over the network. The credentials include `user-name` and the `password` (given by the bank or chosen by the user). The entered credentials are matched with the bank's database, and if found correct, the identity is confirmed. Since, they are open (exposed to view), uncontrolled and unsupervised, they pose several security challenges. Banks are shown to be reluctant replacing completely these schemes with the newer ones because there are no extensive data on their security.

[1] Department of computer Science and Information Engineering (DISI), University of Trento, Via Sommarive 5, Povo, Trento, Email: {attaullah.buriro, sandeep.gupta, bruno.crispo}@unitn.it
[2] DistriNET, KULeuven, Belgium, bruno.crispo@cs.kuleuven.be
[3] https://thefinancialbrand.com/62013/mobile-online-banking-payments-billpay/

This paper proposes motion-assisted touch-typing biometrics - a method to overcome the limitations of PINs/password, for the users of mobile banking apps. The scheme leverages two common human behaviors, i.e., touch-typing and phone-movements by the user. It identifies a user based on the timing differences of the entered strokes and the phone-movements made during the period of text entry. The user is authenticated on the basis of what and how she entered the text. In the case an adversary finds what is being entered, the access will still be denied because of the presence of the two invisible and inherently secure behaviors, i.e., touch-type timing differences and phone-movement. What we propose is also a effortless way to adopt behavioral biometrics, which not only complements the existing traditional methods but also keeps collecting data and security incidents with respect to time to evaluate dynamically the use of behavioral biometric only or both. We did not have time in this paper to collect historical data to say something specific about the choice, but the scheme offers gradual enrollment strategy.

The proposed scheme is fully transparent as it does not require any additional input from a user besides entering the credentials that makes it not only more usable but augments an additional layer enhancing the security of PINs/passwords as mimicking the person-specific movements are extremely onerous. Our scheme utilizes the built-in hardware, i.e., 3-dimensional sensors, to register user-generated phone-movements, and touchscreen to obtain touch-strokes. The sensors are started on the first touch and stopped on the last (the 8^{th}). We evaluated our scheme on our collected dataset of 95 users by applying multi-class classification approach replicating the banking scenario. The main contributions of this paper are listed below:

- The proposal of a secure and usable behavioral-biometric-based authentication solution, for mobile banking. The scheme contemplates the touch-strokes timing-differences and the phone-movements during the process of entering PIN/password.
- Proof-of-concept prototype Android application of a proposed scheme for smartphones.
- Collection and sharing (in the due course of time) of the collected dataset of 95 users.

2 Related Work

Since the behavioral patterns can be collected unobtrusively, behavioral-biometric-based schemes are widely being researched for smartphone user authentication, these days. The search of new human behaviors, profiled through mobile sensors, have gained significant focus these days. Among all the researched schemes, i.e., the way a user walks (gait) [NWB12, De10] and they way a user types/enters any text (touchstroke)[Gi14, Bu15b, Bu16] are very popular.

Our scheme is a bimodal system which leverages the timing-differences from the entered 8-digit secret and the phone-movements while the user enters the text to login to the banking

app, we compare our work with the closely related work proposed in the literature, i.e., [Gi14, Bu15b, Bu16] .

Giuffrida et al., [Gi14], proposed sensor enhanced keystroke based scheme for user authentication on Android smartphones. They reported an Equal Error Rate (EER) of 4.97% and 0.08% on fixed-text passwords (keystroke) and on sensory data, respectively, on a dataset of 20 users. Later, Buriro et. al., [Bu15b, Bu16] modeled sensory readings as hold behavior and introduced free-text secret the user needs to enter or writes on the touchscreen. They reported 1% EER on a dataset of 12 users for touch-typing [Bu15b] and $\approx$ 95% TAR at 3.1% False Acceptance Rate (FAR) on the dataset of 30 users [Bu16].

Our scheme is different from the previously proposed schemes in at least two ways: (i) all these papers performed in-lab supervised experiments and their analysis was based on a small number of users, i.e., just 12 [Bu15b], 20 [Gi14], and 30 [Bu16]. We evaluated our scheme on a comparatively larger dataset of 95 users collected in the wild. Since the number of users in previous studies was less and data was collected in the lab settings, it is difficult to examine how their achieved error would have varied if the number of users was more and data was collected in the wild. (ii) All of the papers evaluated their data either using one class or binary class classification [Bu15b] - replicating authentication on their smartphones [Bu16, Si15], but we have evaluated our data by applying multi-class classification replicating server based remote client authentication.

3 Motion-based Touch-typing Biometrics

To perform an online transaction, the user is required to login to the banking app, which is generally performed by entering the credentials i.e., email, customer-id, and 8-digit PIN/password. Banking server matches the credentials and decides accordingly. Hence, the user is authenticated on the basis of entered text and one who enters the correct pre-stored credentials is treated as the genuine customer/user. Since the password is vulnerable to spoofing, this mechanism poses a threat to the customer privacy.

Our scheme authenticate the user based on what and how she enters the text. Our scheme computes the key-hold and inter-stroke timings from the entered 8-digit secret and extracts the statistical features from different 3-dimensional sensors, for the entire duration of input, to profile the genuine user. In this way, it provides both usability (because the authentication mechanism is hidden from the user/customer), and security (because it is very difficult to impersonate the two inherently secure invisible human behaviors). Thus, the attacker needs to successfully mimic both invisible and person-specific characteristics to get access.

Figure 1 illustrates our approach. In enrollment phase, the banking server collects all the required features from the entered text and phone-movements to form a feature vector. Then it applies feature selection scheme to find out the most productive subset and calls it final feature. This final feature vector is saved in the bank's database under a particular label (i.e., user id, etc.)

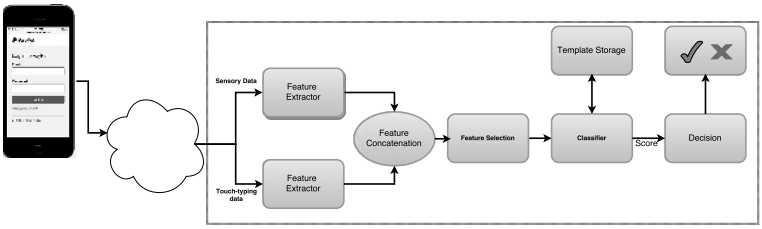


Fig. 1: Model diagram

In verification phase, the user is required to enter the label and the 8-digit password. The banking server picks the earlier pre-selected features from the features of the entered sample, and forms the query feature vector. Later, it compares this feature vector with all the pre-stored feature vectors under that label to find similarity, and authenticates the user, accordingly.

It is our assumption that the users would happily provide these much number of samples for enrollment in final systems, we evaluated our approach for different number of samples, i.e., 5, 10 and 15. However, we consider these samples too few to train the advance machine learning algorithms, i.e., deep learning, we chose simple but effective classifiers, which can perform pretty well even on less training-samples, for our evaluation.

4 Evaluation

4.1 Dataset

We collaborated with "UBERTESTERS a crowd sourcing platform to test the application, involving 95 users. We prototyped an Android application, namely, *PIN&WIN* to collect data. Our application can be installed on any Android device running 4.4.x OS or higher. We setup a web page with the complete explanation of *PIN&WIN*, i.e., the user consent, the procedure to install/uninstall the application. The testers had to agree to the consent form in order to download the app and to participate to the experiment. Then, they had to install the application, answer to the demographic questions, enter $8 - digit$ touch-types and keep the application running for at least 3 days. *PIN&WIN* required user's interaction in 3 sessions in 3 days. *PIN&WIN* required 30-minutes of user interaction on the first day, after installation, and 15 minutes of interaction on the following two days. In this manner, each user had to test the application for 1 hour, however, they needed to keep the application installed for $3 - days$. We collected 30 samples from each participant (in total 2850).

Tab. 1: User demographics (M = Male, F = Female, R = Right, L = Left)

Information	Description	Information	Description
No. of Users	95	Gender	75(m), 20(f)
Sample Size	2, 850 (30 X 95)	Password	8-digit
Devices	Android Smartphones with atleast `4.4.x` version	Handedness	89(R), 6(L)
No. of Sessions	3	Age Groups	90 (20 – 40), 5 (41 – 60)

4.2 Features Extraction & Selection

Our solution leverages all the 3-dimensional sensors i.e., the accelerometer, the orientation, the gravity sensor, the magnetometer and the gyroscope besides the touchscreen. Additionally, it also derives two other sensory readings by applying two filters, i.e., Low-Pass Filter (LPF) and high-Pass Filter (HPF). The value of $\alpha = 0.1$ was computed dynamically[4] to apply to these filters [Bu17]. Our solution leverages sensory readings from all 7 (3-dimensional) sensors in addition to the touchscreen data.

We gathered 4 datastreams from 3-dimensional sensors. Additionally, we computed 4^{th} dimension for all the sensors, and called it magnitude, like in the previous studies [Bu15b][Zh14][Si15] .

We extracted 4 statistical features, namely mean, standard deviation, skewness, and kurtosis, from every data stream [Bu15b, Bu16, BCZ17]. Data from every sensor was transformed into a 4 by 4 features matrix. In total, we obtained 16 features from all four dimensions of each sensor. So the final feature vector for phone-movement behavior, from 7 physical sensors, becomes 112 features long. Similarly, the touch-typing feature vector is 30 features long extracted from the 8-digit password (similar to [Bu15b]). Hence, the final feature vector after concatenation becomes 142 features long.

The primary purpose of any feature selection scheme is to filter out the redundant and less productive features and feed the classifier with the most productive ones. Additionally, this helps also in decreasing the computational cost, i.e., processing smaller feature vectors would take less time. We applied Information Gain Attribute Evaluator[5](IGAE)- a Weka[6] implemented Information Gain based feature selection scheme. This scheme evaluates the worth of a feature by computing its information gain with respect to the class [BCZ17]. We obtained the threshold for feature selection by dividing the number of users (95) by the total number of features (142). The feature with higher weight was picked for further analysis.

[4] https://developer.android.com/reference/android/hardware/SensorEvent.html
[5] http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html
[6] http://www.cs.waikato.ac.nz/ml/weka/downloading.html

4.3 Classifiers

The classifier selection depends upon various parameters, i.e., data size, nature of the data, training time, etc. Our classification toolbox consists of simple but effective state-of-the-art classifiers: Naive Bayes(NB), NeuralNet(NN), and RF classifiers. All these classifiers are considered useful for smaller datasets and found useful in recent studies. We used PRTools[7], a matlab-based toolbox, for all the adopted classifiers and applied all of them in their default settings.

4.4 Experimental Protocol

As we collected 30 observations from each user, we picked first 5, 10 and 15 training samples for simultaneous training of the classifier and used the remaining samples for testing. The training with prior samples looks justified because after repetition the behavior becomes consistent and might show some biased results, i.e, training with prior samples and testing with remaining samples provides comparatively less accuracy.

4.5 Results

We report our obtained results in terms of True Acceptance Rate (TAR), False Reject Rate (FRR), False Acceptance Rate (FAR), True Reject Rate (TRR) and Receiver Operating Characteristics (ROC) curves. In particular, TAR, FAR, FRR and TRR can be defined as the fraction of the genuine samples correctly classified as genuine, the impostor samples incorrectly classified as genuine, the genuine samples incorrectly classified as impostors, and the impostor samples correctly classified as impostor, respectively. Since the FRR and TRR can be estimated by computing $1 - TAR$ and $1 - FAR$, respectively, we show TAR and FAR to avoid redundancy.

In Table 2, we show the TAR and FAR of our chosen classifiers on full and IGAE feature sets. It is evident that the TAR of all the classifiers increased on IGAE features, i.e., for RF classifier, it was 80.51% on full features and it increased upto 89.09% on selected features, for 5 training samples. Similarly, the TAR improved, significantly, as the number of training samples increased, i.e., from 80.51% to 89.09%, from 87.87% to 95.15%, and from 91.79% to 96.00%, for 5, 10, and 15 samples, respectively. The maximum TAR obtained by RF classifier is 96% on 15 training samples.

ROC curves are typically plotted between TAR on the y-axis and FAR on the x-axis. The curve starts from (0,0) and ends at (1,1) coordinates. The curve closer to (0,1) shows the better performance. We show an average ROC of all the users obtained through Vertical

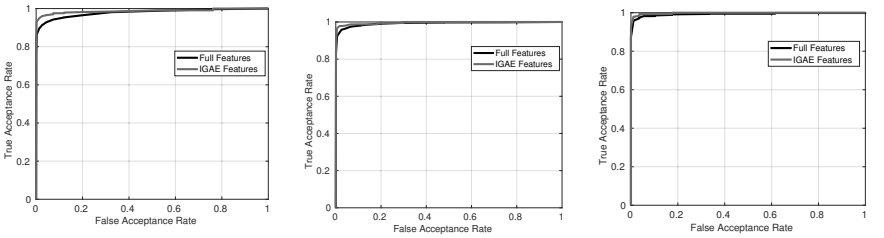
[7] http://prtools.org/

Averaging (VA)[Fa04] in Figure 2. In this averaging, the averages of the TAR rates is plotted against the researcher-defined fixed FAR. Due to the space limitations, we illustrate ROC curves for best performing classifier, i.e., for RF.

RF classifier outperformed both NB and NN classifier because of its ability to reduce the variances and its most unlikeliness of overfitting. NB classifier requires Gaussian distributed data, which might not be true in the dataset, hence it failed to address the problem of concept-drift. The NN classifier failed because of the limited number of training samples. It generally requires higher number of training samples to learn well.

Tab. 2: Results of different classifiers (averaged over all 95 users) on full and IGAE features.

	5				*10*				*15*			
	Full		*IGAE*		*Full*		*IGAE*		*Full*		*IGAE*	
Classifiers	TAR	FAR	TAR	FAR	TAR	FAR	TAR	FAR	TAR	FAR	TAR	FAR
NB	72.72	0.24	79.16	0.19	83.66	0.12	85.11	0.11	87.58	0.07	86.88	0.07
NN	57.81	0.37	77.26	0.20	63.61	0.27	84.51	0.11	70.53	0.16	85.89	0.08
RF	**80.51**	**0.17**	**89.09**	**0.09**	**87.87**	**0.09**	**95.19**	**0.04**	**91.79**	**0.04**	**96.00**	**0.01**



(a) 5 training samples (b) 10 training samples (c) 15 training sample

Fig. 2: ROC curves of RF classifier for (a) 5, (b) 10, and (c) 15 training sample scenarios.

5 Conclusion & Future Work

We have proposed a simple, effective and user-friendly, behavioral biometric-based remote user authentication solution for financial sector. The paper targets the users of mobile banking apps and helps the bank server in identifying the genuine user from the timing-differences of the entered strokes and the movements the user makes while entering the 8-digit secret. Our schemes is user-friendly, as it does not require any extra action for authentication. The transparent additional security layer based on phone motion enhances the security of the scheme, as mimicking simultaneously the two invisible and inherently secure human behaviors is very difficult, if not impossible.

We tried three different classification techniques and RF outperformed the other two. With RF as classifier, we obtained as high as 96% TAR on 15 training samples.

As some papers show [Bu16, Bu15a, Si15] that the behavioral patterns vary in different situations, so it will be interesting to test the scheme in different situations. We have already prototyped the proof-the-concept app based on our findings, however, its evaluation in terms of usability, and robustness against attacks, is a subject of future work. Additionally, its performance evaluation in terms of power consumption, computational constraints, i.e., CPU and memory overhead, and the sample acquisition time and decision time will be investigated as well.

References

[BCZ17] Buriro, A.; Crispo, B.; Zhauniarovich, Y.: Please Hold On: Unobtrusive User Authentication using Smartphone's built-in Sensors. In: IEEE International Conference on Identity, Security and Behavior Analysis (ISBA-2017). 2017.

[Bu15a] Buriro, A.; Crispo, B.; Del Frari, F.; Klardie, J.; Wrona, K.: Itsme: Multi-modal and unobtrusive behavioural user authentication for smartphones. In: International Conference on Passwords. Springer, pp. 45–61, 2015.

[Bu15b] Buriro, A.; Crispo, B.; Del Frari, F.; Wrona, K.: Touchstroke: Smartphone User Authentication Based on Touch-Typing Biometrics. In: proceedings of the New Trends in Image Analysis and Processing– ICIAP 2015 Workshops. Springer, pp. 27–34, 2015.

[Bu16] Buriro, A.; Crispo, B.; Del Frari, F.; Wrona, K.: Hold and Sign: A Novel Behavioral Biometrics for Smartphone User Authentication. In: IEEE Security and Privacy Workshops (SPW). pp. 276–285, 2016.

[Bu17] Buriro, A.: Behavioral Biometrics for Smartphone User Authentication. PhD thesis, University of Trento, 2017.

[De10] Derawi, M.O.; Nickel, C.; Bours, P.; Busch, C.: Unobtrusive user-authentication on mobile phones using biometric gait recognition. In: IEEE 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). pp. 306–311, 2010.

[Fa04] Fawcett, Tom: ROC graphs: Notes and practical considerations for researchers. Machine learning, 31(1):1–38, 2004.

[Gi14] Giuffrida, C.; Majdanik, K.; Conti, M.; Bos, H.: I sensed it was you: authenticating mobile users with sensor-enhanced keystroke dynamics. In: International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, pp. 92–111, 2014.

[NWB12] Nickel, C.; Wirtl, T.; Busch, C.: Authentication of smartphone users based on the way they walk using k-NN algorithm. In: IEEE 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). pp. 16–20, 2012.

[Si15] Sitova, Z.; Sedenka, J.; Yang, Q.; Peng, G.; Zhou, G.; Gasti, P.; Balagani, K.: HMOG: A New Biometric Modality for Continuous Authentication of Smartphone Users. arXiv preprint arXiv:1501.01199, 2015.

[Zh14] Zheng, N.; Bai, K.; Huang, H.; Wang, H.: You are how you touch: User verification on smartphones via tapping behaviors. In: IEEE International Conference on Network Protocols (ICNP). pp. 221–232, 2014.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 227

Steady-State Visual Evoked Potentials for EEG-Based Biometric Identification

Emanuela Piciucco[1], Emanuele Maiorana[1], Owen Falzon[2], Kenneth P. Camilleri[2,3], Patrizio Campisi[1]

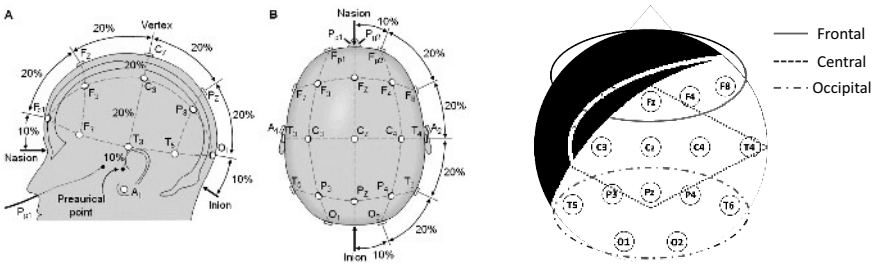
Abstract: In this paper we propose a biometric recognition system based on steady-state visual evoked potentials (SSVEPs), exploiting brain signals elicited by repetitive stimuli having a constant frequency as identifiers. EEG responses to SSVEP stimuli flickering at different frequencies are recorded, and both mel-frequency cepstral coefficients (MFCCs) and autoregressive (AR) reflection coefficients are used as discriminative features of the enrolled users. An analysis of the permanence across time of the brain response to SSVEP stimuli is also performed, by exploiting EEG data acquired in sessions disjoint in time. The employed database is composed by EEG recordings taken from 25 healthy subjects during two different sessions with 15 day average distance between them. The results show that good recognition performance and a high level of permanence can be reached exploiting the proposed method.

Keywords: EEG Recognition, SSVEP, Biometrics.

1 Introduction

Brain signals have been deeply investigated and exploited for medical and brain-computer interface (BCI) purposes since the beginning of the twenty-first century [Ba99]. In recent years, the interest in using such physiological characteristic also for biometric recognition is rapidly increasing. Many studies in such research field have in fact been focused on the use of electroencephalography (EEG) signals, showing that the brain response to specific tasks can be exploited to extract discriminative features able to guarantee high levels of recognition accuracy [CLR14]. The reason for the interest in using EEG data for biometric purposes is linked to some advantages the aforementioned signals possess, compared to other traditional biometric identifiers: universality is in fact guaranteed, and robustness to spoofing attacks and privacy compliance can be easily achieved. In the context of biometric recognition, EEG signals can be recorded as a response to different kinds of stimuli. Specifically, brain signals can be acquired when visual stimuli are presented, that is, when visual evoked potentials (VEPs) are elicited [DMC16, YSL13], or alternatively as a response to tasks such as imagined body movement or speech [MM07, BK10], or while the involved subject is in resting state conditions [NWS07]. In this paper, we propose an EEG-based biometric recognition system where discriminative features are extracted from steady-state visual evoked potentials (SSVEPs). SSVEPs are a particular kind of VEPs that consist of stationary periodic oscillations observed in brain activity as response to a repetitive visual stimulus in the range of 4 Hz to 60 Hz. When an individual focuses his attention on a flickering stimulus within this frequency range, typically presented on an LED setup or LCD display, an increased oscillatory activity, with spectral

[1] Section of Applied Electronics, Department of Engineering, Rome Tre University, 00146 Roma, Italy {emanuela.piciucco, emanuele.maiorana, patrizio.campisi}@uniroma3.it
[2] Centre for Biomedical Cybernetics, University of Malta, Msida 2080, Malta, {owen.falzon, kenneth.camilleri}@um.edn.mt
[3] Faculty of Engineering, University of Malta, Msida 2080, Malta, kenneth.camilleri @um.edu.mt



(a) (b)

Fig. 1: (a) Montage of electrodes used during the acquisition stage. (b) Brain regions.

peaks at the stimulus frequency and its harmonics, can be observed in brain signals [RS98]. SSVEPs exhibit a high signal-to-noise ratio and a stable spectrum, properties which have led to their widespread use for the investigation of cognitive processes such as visual attention and working memory, and clinical conditions such as schizophrenia, autism and epilepsy [Vi10]. These characteristics have also led SSVEPs to being widely adopted in BCI systems, that is, systems allowing an individual to communicate or control equipments solely through their brain activity [Zh10]. The consistent, rapid and prominent response of SSVEPs also makes these signals particularly appealing for EEG-based biometric applications. In contrast with their use in BCI systems, where the primary aim is distinguishing between different visual targets for a given individual, in a biometric system the main challenge lies in identifying features that are sufficiently distinct across individuals, whilst ensuring their stability across multiple recording sessions of the same subject [MLRC16]. The use of SSVEP in biometric applications has been so far investigated only in [Ph16] and [Fa17]. In [Ph16], an analysis based on the peak magnitude and frequency of the short-term Fourier transform has been exploited to identify five users, whose signals have been recorded during a single acquisition session. In [Fa17], the performance of SSVEPs has been assessed for the identification of eight individuals across three recording sessions. Feature vectors consisting of the normalised magnitude responses at a number of stimulus frequencies and their harmonics are computed for each participant. The results obtained indicate that SSVEPs can yield features that are distinct enough between individuals whilst also being sufficiently consistent across multiple sessions for the same individual.

In this paper, a novel approach for EEG recognition based on SSVEPs is proposed. Being the issue of permanence across time of paramount importance for real-life applications of EEG-based biometric systems, the stability of SSVEPs is also specifically addressed. The paper is structured as follows. Section 2 gives an overview of the employed acquisition protocol and the tools used to acquire EEG data. Section 3 describes the proposed biometric recognition system, while the achieved performance and permanence results are reported in Section 4. Some conclusions are eventually drawn in Section 5.

2 Employed Acquisition Protocol

In our work, EEG signals from $U = 25$ healthy volunteers are recorded and used for experimental tests. The device employed to elicit SSVEPs consists of a square array of 9 green leds, whose flickering frequency can be manually tuned. Four different elicitation

frequencies are exploited, namely $f_S \in \mathscr{F}_S = \{6, 12, 18, 24\}$ Hz. During each EEG data acquisition, subjects were comfortably seated on a chair in a dimly lit room, and asked to concentrate on the flickering target for one minute for each considered frequency. The involved subjects were asked to perform the proposed experiment during two temporally separated sessions, referred in the following as S1 and S2. The second session S2 is carried out after an average temporal distance of 15 days from the first session. EEG signals are acquired using a GALILEO BE Light amplifier operating at a sampling rate of 256 Hz. Brain activity is recorded from 19 electrodes placed on the scalp according to the 10-20 international system, as shown in Fig. 1.(a), with potentials referred to an electrode placed at the middle of the central region. At the beginning of each acquisition, the electrical impedance between each electrode and the scalp is kept under $30k\Omega$ using conductive gel. The recorded EEG signals are later preprocessed in order to remove noise and improve signal-to-noise (SNR) ratio, before distinctive features are first extracted and then matched for recognition purposes, as described in Section 3.

3 Employed SSVEP-based Recognition System

The preprocessing applied to the acquired EEG signals is described in Section 3.1. The features employed to represent the collected data are introduced in Section 3.2, while Section 3.3 describes the matching procedure employed in the considered identification system.

3.1 Preprocessing

In order to improve the quality of the acquired EEG signals, a spatial filter, namely a common average referencing (CAR) filter, is first applied to the recorded data. The aim of such filter is to reduce artifacts related to inappropriate reference choices in monopolar recordings [SA15] or unexpected reference variations. Having indicated as $\mathbf{v}_m^{(u)}$, with $u = 1, ..., U$ and $m = 1, ..., M$, the u-th user's potential between the m-th electrode and the reference electrode, filtered data are obtained by computing the difference between the considered EEG signal and the mean of the entire electrode montage:

$$\mathbf{c}_m^{(u)} = \mathbf{v}_m^{(u)} - \frac{1}{M} \sum_{m=1}^{M} \mathbf{v}_m^{(u)} \tag{1}$$

A band-pass filtering is then performed on the CAR-filtered signals. Specifically, since EEG data are characterized by a frequency spectrum with significant elements mainly below 40 Hz, the signals are filtered in the $[0.5, 40]$ Hz band. In order to analyze the brain response behavior, different combinations of the subbands related to the main brain rhythms, that is Delta (δ, $[0.5-4]$ Hz), Theta (θ, $[4-8]$ Hz), Alpha (α, $[8-14]$ Hz), Beta (β, $[13-30]$ Hz) and Gamma (γ, over 30 Hz) are also considered in the performed experimental tests when defining the applied band-pass filter. The obtained data are then downsampled at 128 Hz when the frequency interval of interest comprises the γ subband, otherwise the signals are downsampled at 64 Hz. The so-obtained data are then segmented into R consecutive overlapping frames $\mathbf{y}_m^{(u,r)}$, $r = 1, \ldots, R$, lasting $D = 5$ s with a normalized overlapping factor of $O = 75\%$ between each frame and the previous one.

3.2 Feature Extraction

After EEG data have been preprocessed, discriminative features are evaluated to generate a template from each user u's recording. In this work we exploit two different representations, namely mel-frequency cepstral coefficients (MFCCs) and auto-regressive (AR) coefficients, respectively detailed in Sections 3.2.1 and 3.2.2.

3.2.1 Mel Frequency Cepstral Coefficients (MFCCs)

MFCCs are a parametric representation of the signal based on the Fourier spectrum, widely used in speech-based biometric systems [GFK05] and recently applied to EEG data [Ng12] too. The following steps detail the processing carried out for MFCCs extraction:

1. **power spectral estimate**: the power spectral density (PSD) $\mathbf{Y}_m^{(u,r)}$ of each signal $\mathbf{y}_m^{(u,r)}$, $m = 1,...,M$ and $r = 1,\ldots,R$, is computed through the Welch's averaged modified periodogram approach, using 1-s sliding Hanning windows with 0.5-s overlap;

2. **mel-filter bank processing**: a bank of B mel-filters is used to warp the computed spectrum bins into the mel-scale, defined as:
$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{2}$$
The generated mel-spectrum is indicated in the following as ${}^{MEL}\mathbf{Y}_m^{(u,r)}[b]$, $b = 1,\ldots,B$;

3. **log compression**: the range of the values of the mel-spectrum is reduced through a logarithmic transformation, that is ${}^{LOG}\mathbf{Y}_m^{(u,r)} = ln({}^{MEL}\mathbf{Y}_m^{(u,r)})$;

4. **discrete cosine transform**: MFCCs are computed from the log-compressed mel-spectrum using the discrete cosine transform (DCT):
$$\mathbf{d}_m^{(u,r)}[l] = \sum_{b=1}^{B} {}^{LOG}\mathbf{Y}_m^{(u,r)}[b] \cos\left[l\left(b - \frac{1}{2}\right)\frac{\pi}{B}\right], \quad l = 1,\ldots,L, \quad L < B. \tag{3}$$

In the adopted implementation, $B = 18$ mel-filters are employed, and $L = 12$ DCT coefficients are used to generate the representation of each considered signal. The template associated to the r-th frame of user u's recording, having length $P = M \cdot L$, is eventually obtained by combining the M representations of each channel:
$$\mathbf{f}^{(u,r)} = [\mathbf{d}_1^{(u,r)},\ldots,\mathbf{d}_M^{(u,r)}]. \tag{4}$$

3.2.2 AR Reflection Coefficients

Each EEG frame $\mathbf{y}_m^{(u,r)}$ extracted from the preprocessed signals can be modeled as a realization of an AR stochastic process of order Q, with $Q = 10$ in the adopted implementation. According to such assumption, the available signals can be expressed as:
$$\mathbf{y}_m^{(u,r)}[n] = -\sum_{q=1}^{Q} a_{m,Q,q}^{(u,r)} \mathbf{y}_m^{(u,r)}[n-q] + \mathbf{w}_m^{(u,r)}[n] \tag{5}$$
where $\mathbf{w}_m^{(u,r)}[n]$ is a realization of a white noise process having standard deviation $\sigma_{m,Q}^{(u,r)}$, and $a_{m,Q,q}^{(u,r)}$ are the autoregressive coefficients representing the model. The Yule-Walker equation [Ka88] is used to estimate the Q autoregressive coefficients, employing the recursive Levinson algorithm and introducing the concept of reflection coefficients. In detail, being $a_{m,Q,q}^{(u,r)}$ a generic AR coefficient, we have:
$$\begin{cases} a_{m,Q,q}^{(u,r)} = a_{m,Q-1,q}^{(u,r)} + K_{m,Q}^{(u,r)} \cdot a_{m,Q-1,Q-q}^{(u,r)}, & q = 1,...,Q-1 \\ \sigma_{m,Q}^{(u,r)} = \sigma_{m,Q-1}^{(u,r)} \sqrt{1 - (K_{m,Q}^{(u,r)})^2} \end{cases} \tag{6}$$
where the term $K_{m,Q}^{(u,r)}$ is referred to as reflection coefficient of order Q. In our work, the reflection coefficients are estimated through the Burg method [Ka88], and employed as

representative features of each user u's EEG data. For the generic r-th frame $\mathbf{y}_m^{(u,r)}$ extracted from the m-th channel of the EEG signal belonging to the user u, we therefore generate a feature vector $\mathbf{K}_m^{(u,r)}$ composed of the Q estimated AR reflection coefficients. The overall template associated to a given frame is obtained by combining the M representations generated for each channel into a single vector having size $P = M \cdot Q$, as:

$$\mathbf{f}^{(u,r)} = [\mathbf{K}_1^{(u,r)}, \ldots, \mathbf{K}_M^{(u,r)}]. \tag{7}$$

3.3 Identification

During the identification stage, the Manhattan (L1) distance is used to evaluate the similarity between features extracted during enrolment, and those obtained from an identification probe. In more detail, having indicated as $\mathbf{f}^{(u,e)}$ the template associated with the e-th frame extracted from user u's enrolment, $e = 1, \ldots, E$, and with $\mathbf{f}^{(x,i)}$ the representation generated from the i-th frame taken from the probe of an unknown subject x, $i = 1, \ldots, I$, the distance between such identification frame and the whole set of enrolment frames is evaluated as:

$$d_i^{(u)} = \min_e \left\{ \sum_{p=1}^{P} \left| \mathbf{f}^{(x,i)}[p] - \mathbf{f}^{(u,e)}[p] \right| \right\}, \tag{8}$$

that is, selecting the minimum among the distances computed between the i-th identification frame and all the recorded enrolment data. A decision $\hat{x}_i = \arg\min_u \{d_i^{(u)}\}$ is then taken for each available identification frame, with the final decision $\hat{x}$ regarding the identity of the presented subject taken according to a majority voting rule, selecting the identity with the highest number of occurrences among the votes $\hat{x}_i$, $i = 1, \ldots, I$.

4 Experimental Results

The aim of the present work is to analyze the recognition performance of an EEG-based recognition system exploiting an SSVEP protocol as stimulus for the involved users, taking into account issues regarding repeatability and stability across time of EEG signals. For this purpose, as remarked in Section 2, the collected database comprises EEG recordings taken, for each user, during two disjoint sessions, separated by an average time distance of 15 days. Data from the first session (S1) are considered as enrolment samples, while testing data are selected from the second session (S2). Comparing EEG samples taken during two distinct sessions allows estimating performance depending only on the peculiar characteristics of subject-specific neural activity. This way, session-specific exogenous conditions, such as the capacitative coupling of electrodes and cables with lights or computer, induction loops between the employed equipment and the body, and so on, cannot affect either inter- and intra-class variability of EEG recordings, as instead it may happen when performing tests by comparing EEG data collected during a single acquisition session [MLRC16].

In order to estimate statistically-significant results, a cross-validation procedure is carried out. Specifically, 30 different runs are performed for each of the scenarios described in the following, with 75% of the frames extracted from S1 employed as enrolment dataset for each considered user, and 75% of the frames generated from S2 randomly selected and employed as testing probes at each run.

Channels	SSVEP	EEG subband					
		$[0.5,40]Hz$	$[0.5,30]Hz$	$[4,40]Hz$	$[4,30]Hz$	$[8,40]Hz$	$[8,30]Hz$
All ($M=19$)	$f_S = 6\ Hz$	70.93	76.67	85.07	86.80	73.60	71.87
	$f_S = 12\ Hz$	92.67	93.73	94.40	92.67	84.80	87.20
	$f_S = 18\ Hz$	**94.53**	90.80	88.27	87.07	84.13	80.00
	$f_S = 24\ Hz$	89.73	88.93	87.87	89.33	85.33	85.73
	$f_S \in \mathscr{F}_S$, feat. fus.	97.47	94.67	**99.73**	98.67	99.33	95.73
	$f_S \in \mathscr{F}_S$, score fus.	95.73	**99.33**	96.27	97.33	97.33	93.33
	$f_S \in \mathscr{F}_S$, dec. fus.	99.47	97.33	99.87	**100.00**	98.87	98.00

Tab. 1: Average correct recognition rate (CRR %) obtained over 30 cross-validation runs, using MFCCs as features. The considered subbands are reported in terms of range of associated frequencies.

Channels	SSVEP	EEG subband					
		$[0.5,40]Hz$	$[0.5,30]Hz$	$[4,40]Hz$	$[4,30]Hz$	$[8,40]Hz$	$[8,30]Hz$
All ($M=19$)	$f_S = 6\ Hz$	72.93	74.67	78.53	79.07	66.93	64.40
	$f_S = 12\ Hz$	88.40	88.93	85.47	82.93	79.47	78.67
	$f_S = 18\ Hz$	93.60	93.47	**94.80**	93.47	86.93	83.87
	$f_S = 24\ Hz$	79.73	79.73	88.80	87.20	82.27	88.80
	$f_S \in \mathscr{F}_S$, feat. fus.	96.27	91.60	**98.93**	96.53	93.87	94.13
	$f_S \in \mathscr{F}_S$, score fus.	94.27	92.27	98.93	**99.33**	96.67	88.27
	$f_S \in \mathscr{F}_S$, dec. fus.	**99.73**	98.80	99.60	97.87	98.27	98.53

Tab. 2: Average correct recognition rate (CRR %) obtained over 30 cross-validation runs, using AR reflection coefficients as features. The considered subbands are reported in terms of range of associated frequencies.

The performance obtained when exploiting the considered elicitation frequencies $f_S \in \mathscr{F}_S = \{6,12,18,24\}\ Hz$, and taking into account all the available channels for template generation ($M = 19$), is reported in terms of average correct recognition rate (CRR) in Tables 1 and 2, respectively for MFCC- and AR-based templates. Besides using the considered stimuli separately, they are also jointly employed by fusing their contributions at:

- **feature level**, by concatenating the templates $\mathbf{f}^{(u,r)}$ generated from the r-th frame of user u's EEG collected at different elicitation frequencies, during both enrolment and identification phases;
- **score level**, summing the distances $d_i^{(u)}$ obtained for each i-th identification frame matched with user u's EEG, for signals collected at different elicitation frequencies;
- **decision level**, adopting a majority voting rule over the final decisions $\hat{x}$ individually taken considering EEG data collected at different elicitation frequencies.

As can be seen, for systems employing a single SSVEP elicitation frequency as stimulus, $f_S = 18\ Hz$ guarantees the best achievable identification rates, with CRR = 94.53% obtained using MFCCs to represent EEG data in the [0.5, 40] Hz subband, and CRR = 94.80% employing AR features estimated from EEG signals in the [4, 40] Hz subband. The considered fusion strategies allow to significantly improve such performance, being able to offer a perfect recognition rate (CRR = 100.00%) when a decision-level fusion is performed on information generated through MFCCs, while CRR = 99.73% when exploiting decision-level fusion with AR features.

Channels	SSVEP fusion	MFCCs			AR reflection coefficients		
		$[0.5,30]Hz$	$[4,40]Hz$	$[4,30]Hz$	$[0.5,40]Hz$	$[4,40]Hz$	$[4,30]Hz$
Frontal	Feature	81.73	85.46	79.60	81.87	75.20	74.00
($M=7$)	Score	**94.13**	90.67	84.67	89.46	**91.20**	82.27
	Decision	87.33	91.06	88.80	89.07	88.13	85.60
Central	Feature	78.13	88.27	90.00	80.40	83.60	84.40
($M=7$)	Score	86.00	90.13	89.60	84.00	84.53	88.13
	Decision	89.33	95.33	**95.87**	86.93	93.60	**94.53**
Occipital	Feature	84.67	85.87	84.40	90.80	86.40	86.27
($M=7$)	Score	74.80	78.13	80.93	77.60	79.47	79.47
	Decision	88.16	86.87	**89.33**	88.13	**90.93**	83.47
$\mathscr{M}$	Feature	92.13	89.60	86.26	91.47	88.13	81.06
($M=5$)	Score	90.40	88.80	86.60	83.60	82.13	72.67
	Decision	**96.00**	94.80	93.73	**91.47**	88.13	84.67

Tab. 3: Average correct recognition rate (CRR %) obtained when different spatial configurations are selected and 30 cross-validation runs are performed.

Given the high accuracy obtained when exploiting all the available 19 channels, further tests are carried out to check whether similar results can be obtained while lowering the number of employed channels. It is worth remarking that minimizing the number of employed electrodes is an issue of paramount importance to reduce user inconvenience. In this regard, Table 3 reports the performance obtained when considering only $M=7$ electrodes placed in either frontal, central and occipital areas, according to the montages shown in Fig. 1.(b), together with the rates obtained with an even smaller set $\mathscr{M}=\{\mathrm{F}_z, \mathrm{C}_z, \mathrm{P}_z, \mathrm{O}_1, \mathrm{O}_2\}$ with $M=5$ electrodes, comprising only midline and occipital channels. Only the recognition rates achieved exploiting all the considered elicitation frequencies through fusion approaches, and taking into account the best-performing subbands according to the results shown in Tables 1 and 2, are reported in Table 3. From the obtained accuracies it can be seen that the central area of the scalp seems guaranteeing the best performance achievable with a reduced number of electrodes, achieving CRR = 95.87% for MFCC and CRR = 94.53% for AR representations, when considering EEG recordings filtered in the $\theta \cup \alpha \cup \beta$ subband. An even better result is obtained when considering only the set $\mathscr{M}$ with $M=5$ in the $\delta \cup \theta \cup \alpha \cup \beta$ subband, for which a CRR = 96.00% is achieved using MFCCs, while AR features provides CRR = 91.47%.

5 Conclusions

This paper evaluates the feasibility of designing an automatic biometric recognition system exploiting EEG signals elicited through protocols generating steady-state visual evoked potentials (SSVEPs). The use of flickering stimuli at specific frequencies and the representation of the acquired EEG data through either MFCC or AR templates, allows achieving high identification rates, thanks to the proved existence of permanent characteristics in SSVEP brain responses across different acquisition sessions. According to the reported experimental tests, the joint use of multiple elicitation frequencies guarantees a notable improvement in recognition rates, thus allowing to reduce the number of electrodes needed during EEG collection, a relevant property to foster the adoption of EEG-based biometric identifiers in practical recognition systems.

References

[Ba99] Başar, E.: Brain Function and Oscillations: Integrative brain function. Neurophysiology and cognitive processes. Springer series in synergetics. Springer, 1999.

[BK10] Brigham, Katharine; Kumar, BVK Vijaya: Subject identification from electroencephalogram (EEG) signals during imagined speech. In: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE Int. Conf. on. IEEE, pp. 1–8, 2010.

[CLR14] Campisi, Patrizio; La Rocca, Daria: Brain waves for automatic biometric-based user recognition. IEEE Trans. on Information Forensics and Security, 9(5):782–800, 2014.

[DMC16] Das, Rig; Maiorana, Emanuele; Campisi, Patrizio: EEG biometrics using visual stimuli: A longitudinal study. IEEE Signal Processing Letters, 23(3):341–345, 2016.

[Fa17] Falzon, Owen; Zerafa, Rosanne; Camilleri, Tracey; Camilleri, Kenneth P.: EEG-Based Biometry Using Steady State Visual Evoked Potentials. In: Proc. of 39th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. IEEE, 2017.

[GFK05] Ganchev, Todor; Fakotakis, Nikos; Kokkinakis, George: Comparative evaluation of various MFCC implementations on the speaker verification task. In: Proceedings of the SPECOM. volume 1, pp. 191–194, 2005.

[Ka88] Kay, Steven M: Modern spectral estimation. Pearson Education India, 1988.

[MLRC16] Maiorana, Emanuele; La Rocca, Daria; Campisi, Patrizio: On the permanence of EEG signals for biometric recognition. IEEE Trans. on Information Forensics and Security, 11(1):163–175, 2016.

[MM07] Marcel, Sebastien; Millán, José del R: Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(4), 2007.

[Ng12] Nguyen, P.; Tran, D.; Huang, X.; Sharma, D.: A proposed feature extraction method for EEG-based person identification. In: Int. Conf. on Artificial Intelligence (ICAI). 2012.

[NWS07] Näpflin, Markus; Wildi, Marc; Sarnthein, Johannes: Test–retest reliability of resting EEG spectra validates a statistical signature of persons. Clinical Neurophysiology, 118(11):2519 – 2524, 2007.

[Ph16] Phothisonothai, Montri: An investigation of using SSVEP for EEG-based user authentication system. 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2015, (December):923–926, 2016.

[RS98] Rager, Günter; Singer, Wolf: The response of cat visual cortex to flicker stimuli of variable frequency. European Journal of Neuroscience, 10(5):1856–1877, 1998.

[SA15] Schwartz, Mark S; Andrasik, Frank: Biofeedback: A practitioner's guide. Guilford Publications, 2015.

[Vi10] Vialatte, François-Benoît; Maurice, Monique; Dauwels, Justin; Cichocki, Andrzej: Steady-state visually evoked potentials: focus on essential paradigms and future perspectives. Progress in neurobiology, 90(4):418–438, 2010.

[YSL13] Yeom, Seul-Ki; Suk, Heung-Il; Lee, Seong-Whan: Person authentication from neural activity of face-specific visual self-representation. Pattern Recognition, 46(4), 2013.

[Zh10] Zhu, Danhua; Bieger, Jordi; Molina, Gary Garcia; Aarts, Ronald M: A survey of stimulation methods used in SSVEP-based BCIs. Comp. Int. and Neuroscience, 2010.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 235

Improving Very Low-Resolution Iris Identification Via Super-Resolution Reconstruction of Local Patches

Fernando Alonso-Fernandez[1], Reuben A. Farrugia[2], Josef Bigun[3]

Abstract: Relaxed acquisition conditions in iris recognition systems have significant effects on the quality and resolution of acquired images, which can severely affect performance if not addressed properly. Here, we evaluate two trained super-resolution algorithms in the context of iris identification. They are based on reconstruction of local image patches, where each patch is reconstructed separately using its own optimal reconstruction function. We employ a database of 1,872 near-infrared iris images (with 163 different identities for identification experiments) and three iris comparators. The trained approaches are substantially superior to bilinear or bicubic interpolations, with one of the comparators providing a Rank-1 performance of ∼88% with images of only 15×15 pixels, and an identification rate of 95% with a hit list size of only 8 identities.

Keywords: Iris, biometrics, super-resolution, low resolution.

1 Introduction

While the literature on image super-resolution is ample, its application to biometrics is relatively recent, with most research concentrated on face reconstruction [Wa14]. However, a number of applications which are becoming ubiquitous, such as surveillance or smartphone biometrics, have the lack of pixel resolution as one of their most evident problems when acquisition is done distantly. One reason of such limited research might be that most super-resolution approaches are general-scene, aimed at producing overall visual enhancement, which does not necessarily correlate with better recognition performance [Ng12]. Thus, adaptation of super-resolution techniques to the particularities of images from a specific biometric modality is needed to achieve a more efficient up-sampling [BK02].

This paper investigates two trained super-resolution approaches based on PCA Eigen transformation (eigen-patches) [AFB15] and Locality-Constrained Iterative Neighbor Embedding (LINE) of local image patches [Ji14] in the context of iris identification. The methods employed make use of coupled dictionaries to learn the mapping relation between low- and high-resolution image pair in order to hallucinate a high-resolution image from the observed low-resolution one. This *learning-based* strategy has the advantage of only needing one low-resolution image as input, and usually allow higher magnification factors than *reconstruction-based* methods, which fuse several low-resolution images into a high-resolution one [PPK03]. Another particularity of the evaluated methods is that they

[1] School of ITE, Halmstad University, Sweden, feralo@hh.se
[2] Department of CCE, University of Malta, Malta, reuben.farrugia@um.edu.mt
[3] School of ITE, Halmstad University, Sweden, josef.bigun@hh.se

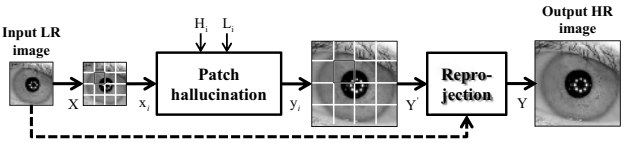


Fig. 1: Block diagram of patch-based hallucination.

use a patch-based approach, where overlapped local image patches are reconstructed separately, and then stitched together. This better represents local details and preserves texture than if reconstruction of the complete image was done at a time, since each patch has its own optimal reconstruction function. In our experiments, we employ the CASIA-IrisV3-Interval database [CA] of NIR iris images, with low-resolution images having a size of only 15×15 pixels. Identification experiments are conducted with three iris comparators based on 1D Log-Gabor filters (LG) [Ma03], SIFT key-points [Lo04], 5 and local intensity variations of iris textures (CR) [RU10]. LG and CR exploit texture information globally (across the entire image), while SIFT exploits local features in discrete key-points. Thus, one motivation is to employ features that are diverse in nature. Despite the patch-based approaches used are not new [AFB15, Ji14], we contribute with its evaluation in the context of iris identification, and particularly with the application of these three iris comparators to the reconstructed images. Reported results show the superiority of the two trained reconstruction approaches w.r.t. bicubic or bilinear interpolations, with an impressive Rank-1 performance of $\sim$88% with the LG comparator under such very low resolution.

2 Reconstruction of Low Resolution Iris Images

Given an input low resolution (LR) image $\mathbf{X}$, the goal is to reconstruct its high resolution (HR) counterpart $\mathbf{Y}$. The LR image can be modeled as the HR image manipulated by blurring (B), warping (W) and down-sampling (D) as $\overline{X} = DBW\overline{Y} + \overline{n}$ (where $\overline{n}$ represents additive noise). For simplicity, W and $\overline{n}$ are usually omitted, leading to $\overline{X} = DB\overline{Y}$. In local patch-based methods (Figure 1), LR images are first separated into $N = N_v \times N_h$ overlapping patches $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ according to a predefined patch size and overlap pixels (N_v and N_h are the vertical and horizontal number of patches). Since we will consider square images, we assume that $N_v = N_h$. Two super sets of basis patches $\mathbf{H}_i$ and $\mathbf{L}_i$ are computed for each patch $\mathbf{x}_i$ from collocated patches of a training database of M high resolution images $\{\mathbf{H}\}$. Super set $\mathbf{H}_i = \left\{\mathbf{h}_i^1, \mathbf{h}_i^2, \cdots, \mathbf{h}_i^M\right\}$ is obtained from collocated patches of $\{\mathbf{H}\}$. By degradation (low-pass filtering and down-sampling), a low-resolution database $\{\mathbf{L}\}$ is obtained from $\{\mathbf{H}\}$, and the other super set $\mathbf{L}_i = \left\{\mathbf{l}_i^1, \mathbf{l}_i^2, \cdots, \mathbf{l}_i^M\right\}$ is obtained similarly from $\{\mathbf{L}\}$. Each individual LR patch $\mathbf{x}_i$ is then hallucinated using the dictionaries $\mathbf{H}_i$ and $\mathbf{L}_i$, producing the corresponding HR patch $\mathbf{y}_i$.

Eigen-Patch Reconstruction Method (PCA). This method is described in [AFB15], which is based on the algorithm for face images of [CC14]. Here, a PCA eigen- transformation is conducted in the set of LR basis patches $\mathbf{L}_i$. Given an input LR patch $\mathbf{x}_i$, it is then projected onto the eigenpatches of $\mathbf{L}_i$, obtaining the optimal reconstruction weights $\mathbf{c}_i = \left\{c_i^1, c_i^2, \cdots, c_i^M\right\}$ of $\mathbf{x}_i$ w.r.t. $\mathbf{L}_i$. The reconstruction weights are then carried on to

weight the HR basis set, and the HR patch is super-resolved as $\mathbf{y}_i = \mathbf{H}_i \mathbf{c}_i^T$. Finally, once the overlapping reconstructed patches $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N\}$ are obtained, they are stitched together by averaging, resulting in the preliminary reconstructed HR image $\mathbf{Y}'$.

Locality-Constrained Iterative Neighbour Embedding Method (LINE). This is based on the algorithm for face images of [Ji14]. Instead of using all entries of the training dictionary to estimate the reconstruction weights, a set of $K < M$ entries is used. Using all entries can result in over-smooth reconstructed images which lacks important texture information, which is essential for iris. Given a LR patch $\mathbf{x}_i$, a first estimate of the HR patch $\mathbf{v}_{i,0}$ is initialized by bicubic up-scaling. Then, an iterative loop indexed by $j \in [0, J-1]$ is started. For every iteration, the support $\mathbf{s}$ of $\mathbf{H}_i$ that minimizes the distance $\mathbf{d} = ||\mathbf{v}_{i,j} - \mathbf{H}_i(\mathbf{s})||_2^2$ is computed using K-nearest neighbours. The combination weights are then derived using

$$\mathbf{w}_{i,j}^* = \arg\min_{\mathbf{w}_{i,j}^*} \left(\left\| \mathbf{x}_i - \mathbf{L}_i(\mathbf{s})\, \mathbf{w}_{i,j}^* \right\|_2^2 + \tau \left\| \mathbf{d}(\mathbf{s}) \odot \mathbf{w}_{i,j}^* \right\|_2^2 \right) \tag{1}$$

where τ is a regularization parameter. Operator $\odot$ (element-wise multiplication) is used to penalize the reconstruction weights with the distances between $\mathbf{v}_{i,j}$ and its closest neighbors in the training dictionary $\mathbf{H}_i$. This optimization problem can be solved by an analytic solution [Ji14]. The estimated HR patch is then updated using $\mathbf{v}_{i,j+1} = \mathbf{H}_i(\mathbf{s})\mathbf{w}_{i,j}^\star$ and the loop is repeated. The final estimate of the HR patch is then derived using $\mathbf{y}_i = \mathbf{v}_{i,J}$. We employ τ=$1e^{-5}$ and J=4 [Ji14]. Contrarily to the PCA method, where reconstruction weights are obtained in the LR manifold and then simply transferred to the HR manifold, note that Equation 1 jointly considers the LR manifold (via $\mathbf{x}_i$, $\mathbf{L}_i(\mathbf{s})$) and the HR counterpart (via $\mathbf{d}(\mathbf{s})$) during the reconstruction. In addition, reconstruction starts in the HR manifold, which is not affected by the degradation process, and computation of the K nearest neighbors employed for reconstruction is done in this manifold as well.

Image Reprojection. Inspired by [AFB15], we incorporate a re-projection step to $\mathbf{Y}'$ to reduce artifacts and make the output image $\mathbf{Y}$ more similar to the input image $\mathbf{X}$. The image $\mathbf{Y}'$ is re-projected to $\mathbf{X}$ via $\mathbf{Y}^{t+1} = \mathbf{Y}^t - \upsilon U\left(B\left(DB\mathbf{Y}^t - \mathbf{X}\right)\right)$ where U is the up-sampling matrix. The process stops when $|\mathbf{Y}^{t+1} - \mathbf{Y}^t| \leq \varepsilon$. We use υ=0.02 and $\varepsilon = 10^{-5}$ [AFB15].

3 Experimental Framework

We use CASIA Interval v3 iris database [CA]. It has 2,655 NIR images of 280×320 pixels from 249 contributors captured with a close-up camera. Manual annotation is available, including iris circles and noise mask (Figure 2) [AB15, Ho14], which is used as input for our experiments. All images are resized by bicubic interpolation to have the same sclera radius (R=105, average of the database given by the ground-truth). Then, images are aligned by extracting a region of 231×231 around the pupil center (corresponding to $\sim$1.1×R). If extraction is not possible (for example if the eye is close to a boundary), the image is discarded. After this procedure, 1,872 images remain, which are then divided into two sets, a training set with images from the first 116 users (M=925 images) used to train the hallucination methods, and a test set from the remaining 133 users (947 images) for

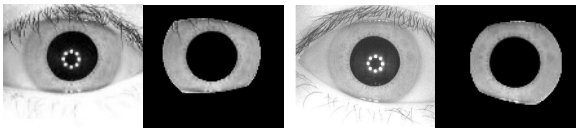


Fig. 2: Example of images of the CASIA Interval v3 database with the annotated circles modeling iris boundaries and eyelids.

validation. We carry out identification experiments with three iris comparators in the test set. From the 133 users, we select those eyes having at least two samples, resulting in 163 different eyes (i.e. identities) and 927 images. The first sample of each eye is considered as enrolment sample, and the remaining 764 samples are used as input for identification. This results in 764×163=124,532 comparisons. Given an input sample, identification is done by outputting the *N* closest identities of the enrolment set. An identification is considered successful if the correct identity is among the *N* outputted ones.

The iris comparators used are based on 1D Log-Gabor filters (LG) [Ma03], SIFT operator [Lo04], and local intensity variations in iris textures (CR) [RU10]. In LG, the iris region is first unwrapped to a normalized rectangle of 20×240 pixels [Da04] and next, a 1D Log-Gabor wavelet is applied plus phase binary quantization to 4 levels. Comparison between binary vectors is done using the normalized Hamming distance [Da04]. In the SIFT method, SIFT key points are directly extracted from the iris region (without unwrapping), and the recognition metric is the number of matched key points, normalized by the average number of detected key-points in the two images under comparison. The CR method starts by unwrapping the iris to a rectangle of 64×512 pixels, and then it traces intensity variations across horizontal stripes of distinct height, encoding the paths where the minimum and maximum grey values of each column occur. The LG implementation is from Libor Masek [Ma03], using its default parameters. The SIFT method uses a free toolkit[3], with adaptations described in [Al09] to remove spurious matchings. The CR algorithm is from the University of Salzburg Iris Toolkit (USIT) [RUW13].

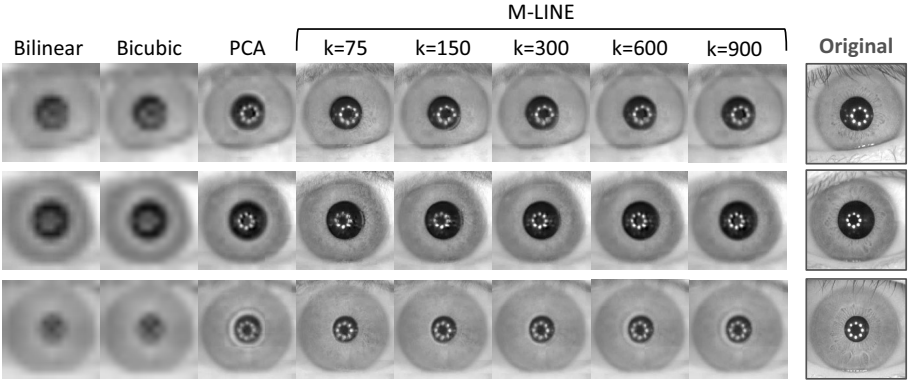


Fig. 3: Resulting hallucinated HR images. The original HR image is also shown.

[3] http://vision.ucla.edu/ vedaldi/code/sift/assets/sift/index.html

4 Results

The two reconstruction methods are evaluated together with bilinear and bicubic interpolations. The 947 validation images are used as HR reference images. They are down-sampled via bicubic interpolation to a size of 15×15, corresponding to a down-sampling factor of 16, and then used as input LR images of the reconstruction methods, from which hallucinated HR images are computed. This simulated down-sampling is the approach followed in most previous studies [Wa14], mainly due to the lack of databases with LR and corresponding HR reference images. In PCA and LINE, we employ a patch size of 1/4 of the LR image size. This is motivated by [AFB15], where better results were obtained with bigger patch sizes. Overlapping between patches is 1/3 of the patch size. We also extract the LG, SIFT and CR features from both the hallucinated HR and the reference HR images. Figure 3 shows some examples of reconstructed images with the different methods tested here. It can be observed that smaller values of K results in sharper reconstructed images, while a bigger K produces blurrier images. This is expected, since a bigger K implies that more patches are being averaged, so the output image patch will be smoother.

The performance of the reconstruction methods is measured by reporting identification experiments using hallucinated images. We do not report other measures traditionally used in super-resolution literature (e.g. PSNR) since the aim of applying these algorithms in biometrics is enhancing recognition performance [Ng12]. Two scenarios are considered: 1) enrolment samples taken from original HR input images, and query samples from hallucinated HR images; and 2) both enrolment and query samples taken from hallucinated HR images. The first case simulates a controlled enrolment scenario, while the second case simulates a totally uncontrolled scenario (albeit for simplicity, both samples have similar resolution). We first test the LINE method using different values of K, from K=75 (small neighbors set) to K=900 (nearly the whole training set). Identification results are given in Figure 4. It can be seen that the preferred neighbor size K is different for each comparator. While LG and CR prefer a bigger set ($K > 300$), SIFT shows better results with a smaller set ($K = 150$). This highlights the need of looking into the performance of individual comparators, rather than into general scene indicators such as PSNR, since the image properties recovered by a particular algorithm may not be relevant for a comparator, even if visual appearance of the reconstructed image can be referred as 'good'.

We then select the best LINE configurations for each comparator, and report identification results together with the other reconstruction methods (Figure 5). Our first observation is the superior performance of PCA and LINE w.r.t. bilinear or bicubic interpolation, highlighting the benefits of trained reconstruction. Also, LINE is superior to PCA in some cases, while in others, both methods show similar performance. In this sense, PCA can be pre-trained in advance using the set $\mathbf{L}_i$ of basis patches, since eigenpatches are the same for any input patch $\mathbf{x}_i$, so higher computational speeds can be expected. LINE on the other hand needs to compute the set of nearest neighbors specific of a particular input patch.

Regarding performance of individual comparators, LG is clearly superior to the others. Rank-1 performance of LG is above 70% (scenario 1) and 84% (scenario 2). Also, an identification rate of 95% with this comparator is obtained for a hit list size of just N=8 (sce-

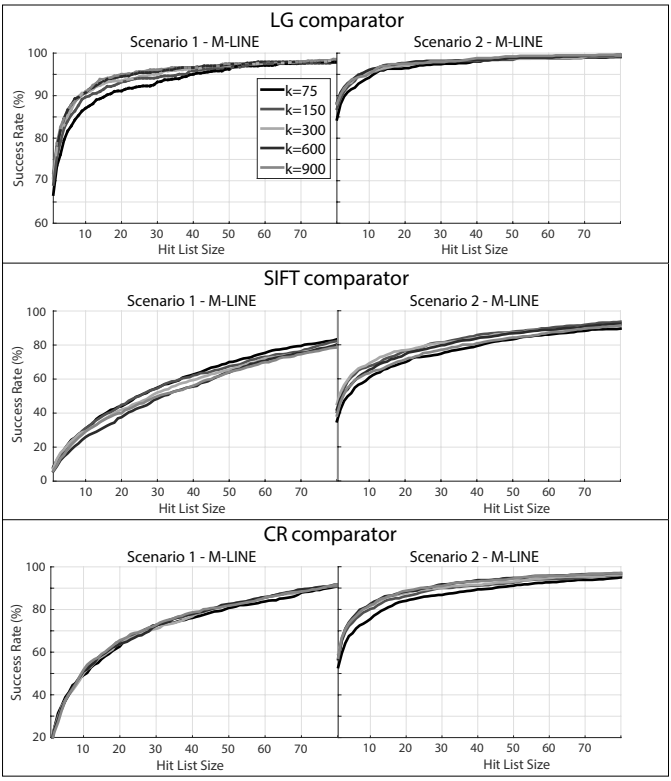


Fig. 4: Identification results (LINE method). Best seen in colour.

nario 2) using LINE. Rank-1 of SIFT is very poor (less than 10% in scenario 1 and ∼40% in scenario 2), while an identification rate of 95% cannot be achieved even if $N > 80$). The CR comparator only does a little bit better than SIFT. It should be noted however that the size of the LR images is very small (15×15). With respect to the two scenarios evaluated, scenario 2 has much better performance. In scenario 2, both enrolment and query images undergo the same down-sampling and reconstruction. It seems that when the two images do not suffer the same degradation process (i.e. scenario 1), they have fairly different feature properties, at least with the features employed here. This result has been observed in previous verification studies [AFB15] as well.

5 Conclusions

While more relaxed acquisition environments are pushing image-based biometrics (e.g. face or iris) towards the use of low resolution imagery, it can pose significant problems in terms of reduced performance if not addressed properly. Here, we apply two trained super-resolution approaches based on PCA transformation [AFB15] and Locality-Constrained Iterative Neighbor Embedding (LINE) of local patches [Ji14] to improve the resolution

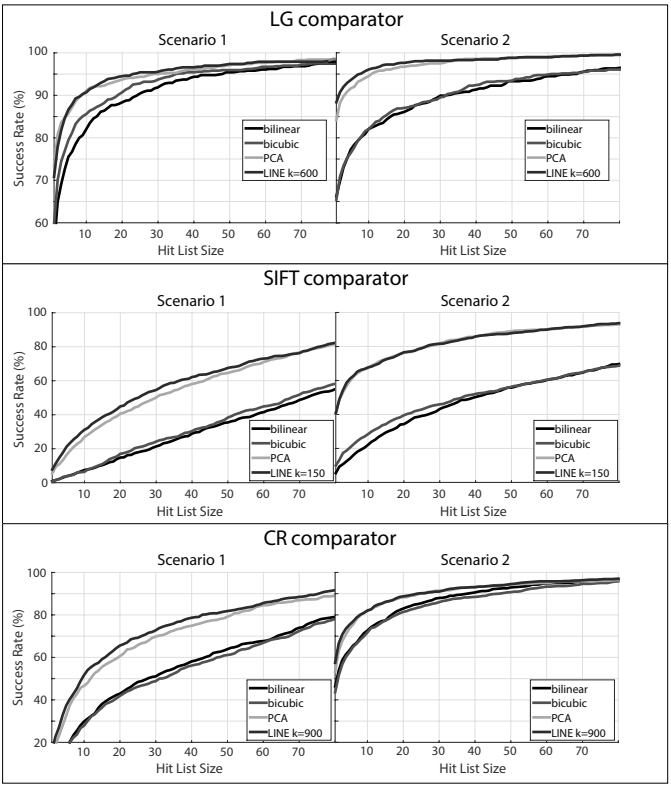


Fig. 5: Identification results of the different image reconstruction methods employed (LINE method: best case according Figure 4 is shown). Best seen in colour.

of iris images under infra-red lightning. We carry out identification experiments on the reconstructed images with three iris comparators based on Log-Gabor wavelets (LG), SIFT keypoints, and local intensity variations of iris textures (CR). Low resolution images are simulated by down-sampling high-resolution irises to a size of just 15×15. Experimental results show a clear superiority of trained approaches under such challenging conditions w.r.t. bilinear or bicubic methods. Even under such low resolution, a Rank-1 performance of $\sim$88% is obtained with one of the comparators (LG), and an identification rate of 95% is obtained with a hit list size of just 8. Another observation is that the LINE method is superior to PCA in some cases, but their performance is in general very similar. This allows computational savings by using PCA, since PCA models are the same for any input image, so they can be trained in advance.

An avenue of improvement is removing the assumption that reconstruction weights are the same in the low- and high-resolution manifolds. While this simplifies the problem, the LR manifold is usually distorted by the one-to-many relationship between LR and HR patches [Wa14]. Another simplification is the assumption of linearity in the combination of

patches from the training dictionary. We will also consider including additional recognition methods [RUW13] and employing imagery in visible range (e.g. smart-phones).

Acknowledgements

Author F. A.-F. thanks the Swedish Research Council for funding his research. Authors acknowledge the CAISR program and the SIDUS-AIR project of the Swedish Knowledge Foundation.

References

[AB15] Alonso-Fernandez, F.; Bigun, J.: NIR & visible-light periocular recognit with Gabor features using frequency-adaptive automatic eye detection. IET Biometrics, 4(2), 2015.

[AFB15] Alonso-Fernandez, F.; Farrugia, R. A.; Bigun, J.: Eigen-Patch Iris Super-Resolution for Iris Recognition Improvement. Proc EUSIPCO, Sep 2015.

[Al09] Alonso-Fernandez, F.; Tome-Gonzalez, P.; Ruiz-Albacete, V.; Ortega-Garcia, J.: Iris Recognition Based on SIFT Features. Proc IEEE BIDS, 2009.

[BK02] Baker, S.; Kanade, T.: Limits on super-resolution and how to break them. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(9):1167–1183, Sep 2002.

[CA] CASIA databases: http://biometrics.idealtest.org/.

[CC14] Chen, Hong-Yuh; Chien, Shao-Yi: Eigen-patch: Position-patch based face hallucination using eigen transformation. Proc IEEE, 2014.

[Da04] Daugman, J.: How iris recognition works. IEEE Trans. on CSVT, 14:21–30, 2004.

[Ho14] Hofbauer, H.; Alonso-Fernandez, F.; Wild, P.; Bigun, J.; Uhl, A.: A Ground Truth for Iris Segmentation. Proc Intl Conf Pattern Recognition, ICPR, 2014.

[Ji14] Jiang, Junjun; Hu, Ruimin; Wang, Zhongyuan; Han, Zhen: Face Super-Resolution via Multilayer Locality-Constrained Iterative Neighbor Embedding and Intermediate Dictionary Learning. IEEE Transactions on Image Processing, 23(10):4220–4231, Oct 2014.

[Lo04] Lowe, D.: Distinctive image features from scale-invariant key points. Intl Journal of Computer Vision, 60(2):91–110, 2004.

[Ma03] Masek, Libor: Recognition of human iris patterns for biometric identification. MSc thesis, School of Computer Science / Software Engineering, Univ Western Australia, 2003.

[Ng12] Nguyen, K.; Sridharan, S.; Denman, S.; Fookes, C.: Feature-domain super-resolution framework for Gabor-based face and iris recognition. Proc CVPR, 2012.

[PPK03] Park, Sung Cheol; Park, Min Kyu; Kang, Moon Gi: Super-resolution image reconstruction: a technical overview. Signal Processing Magazine, IEEE, 20(3):21–36, May 2003.

[RU10] Rathgeb, Christian; Uhl, Andreas: Secure Iris Recognition Based on Local Intensity Variations. Proc ICIAR, 6112:266–275, 2010.

[RUW13] Rathgeb, Christian; Uhl, Andreas; Wild, Peter: Iris Biometrics - From Segmentation to Template Security, volume 59 of Advances in Information Security. Springer, 2013.

[Wa14] Wang, N.; Tao, D.; Gao, X.; Li, X.; Li, J.: A Comprehensive Survey to Face Hallucination. Intl Journal of Computer Vision, 106(1):9–30, 2014.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 243

Deep Quality-informed Score Normalization for Privacy-friendly SpeakerRecognition in unconstrained Environments

Andreas Nautsch[1], Søren Trads Steen[1,2], Christoph Busch[1]

Abstract: In scenarios that are ambitious to protect sensitive data in compliance with privacy regulations, conventional score normalization utilizing large proportions of speaker cohort data is not feasible for existing technology, since the entire cohort data would need to be stored on each mobile device. Hence, in this work we motivate score normalization utilizing deep neural networks. Considering unconstrained environments, a quality-informed scheme is proposed, normalizing scores depending on sample quality estimates in terms of completeness and signal degradation by noise. Utilizing the conventional PLDA score, comparison i-vectors, and corresponding quality vectors, we aim at mimicking cohort based score normalization optimizing the C_{llr}^{min} discrimination criterion.

Examining the I4U data sets for the 2012 NIST SRE, an 8.7% relative gain is yielded in a pooled 55-condition scenario with a corresponding condition-averaged relative gain of 6.2% in terms of C_{llr}^{min}. Robustness analyses towards sensitivity regarding unseen conditions are conducted, i.e. when conditions comprising lower quality samples are not available during training.

Keywords: speaker recognition, score normalization, unconstrained environments, neural networks, deep learning

1 Introduction

Accounting for European data privacy regulations [Eu16], resource limitations of mobile operating scenarios, and technological requirements concerning vast signal quality variations in unconstrained environment speaker recognition, current score normalization schemes are put to its limits. In this paper, we propose a quality-informed score normalization scheme utilizing cohort data for the purpose of training a neural network in order to avoid a distribution of biometric data from cohort subjects, substituting conventional cohort-based score normalization. This study is limited with respect to deeper network architectures and the sensitivity to unseen quality conditions. Comparative experiments to conventional normalization schemes are excluded, since we assume their design to be prohibited due to a restrictive interpretation of §9 in EU regulation 2016/679, i.e. cohort data which is necessary to estimate parameters of zero-norms shall not be distributed. The EU regulation 2016/679 [Eu16, §9] prohibits the processing of biometric data, if not – among others – the biometric subject is giving consent, and the *processing relates to personal data which are manifestly made public by the data subject*. Hence, the distribution and use of cohort data related to other individuals than the biometric subject under processing may

[1] da/sec — Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {andreas.nautsch,christoph.busch}@{crisp-da | h-da}.de
[2] Technical University of Denmark, Denmark, stradssteen@gmail.com

become improper to justify as cohort data would need to be transmitted to the device of any other biometric user for conducting cohort normalization, especially for data deletion.

This paper is organized as follows: Sec. 2 depicts the related work on speaker recognition and neural networks. Sec. 3 depicts the proposed normalization scheme. Experimental evaluations are carried out in Sec. 4, and conclusions are drawn in Sec. 5.

2 Related Work

Recent speaker recognition approaches rely on i-vectors, representing the characteristic speaker offset from an Universal Background Model (UBM), which models the distribution of acoustic features, such as Mel-frequency cepstral coefficients [RQD00]. Thereby, UBM components' mean vectors are concatenated to a *supervector* $\vec{\mu}_{\mathrm{UBM}}$. Speaker supervectors $\vec{s}$ are decomposed by a total variability matrix into a lower-dimensional i-vector $\vec{i}$ as an offset to the UBM supervector $\vec{\mu}_{\mathrm{UBM}}$ [Ke05, De11]. Then, i-vectors are projected onto a spherical space by whitening transform and length normalization [GREW11, BBM13]. State-of-the-art i-vector comparators, e.g. Probabilistic Linear Discriminant Analysis (PLDA) [CL14], conduct a likelihood ratio scoring.

2.1 Conventional Score Normalization Methods

State-of-the-art recognition systems [Va16, Br16] utilize score normalization in order to improve discrimination power on secure operating points by employing statistics from comparisons of the reference against an independent (cohort) data set, referred to as z-norm, from comparisons of the probe against a cohort set, referred to as t-norm, and variations of z- and t-norm, such as the zt-norm, or s-norm, as well as adaptive variations e.g., at- [SR05] and as-norm [Cu11]. Exemplary, in [SR05, Ha13], data of 550, 1039 female, and 435, 680 male speakers is utilized for normalization purposes, respectively, whereas in [Cu11], solely the usage of 348 female and 273 male voice samples is reported. In mobile applications, where no data of the biometric subject should leave the device, the cohort data needs to be present on each mobile device.

2.2 Different Environmental Conditions

Variations in signal quality, i.e. in the probe sample condition, result in different score distributions per condition [Ma13, MSvL15]. While systems are usually calibrated for known scenarios and in fixed-condition environments, calibrating systems well among known as well as unseen conditions is harder, i.e. when facing unconstrained environments.

In this paper, we examine the 55 duration and noise conditions presented in [Na15]. In [Na15], SNR conditions stem from two noise sources: air conditioner (AC) and crowd (CROWD) noise. By degenerating voice samples from the I4U file list [Sa13], combined signal degradation and observation incompleteness (short probe segment duration) effects are simulated, which are expected to represent the most common conditions, cf. Tab. 1.

Tab. 1: *Label scheme for combined duration and noise conditions, cf. [Na15].*

Condition	1	2	3	4	5	6	7	8	9	10	11 … 30	31 … 55
Duration	5 s	10 s	20 s	40 s	full	5 s					10 s … full	5 s … full
Noise	clean					AC						CROWD
SNR						0 dB	5 dB	10 dB	15 dB	20 dB	0 dB … 20 dB	0 dB … 20 dB

2.3 Estimation of Unified Audio Quality Vectors

For the purpose of estimating quality in speaker recognition, unified audio characteristics [Fe12] are utilized. Single multivariate Gaussian models $\Lambda_j \sim \mathcal{N}(\mu_j, \Sigma), j = 1, \ldots, 55$ are trained in original i-vector space for each quality condition as outlined in Tab. 1. The models have condition-dependent mean vectors μ_j and share a full covariance matrix Σ. Class-dependent means are estimated using i-vectors from a respective quality condition and Σ is estimated by pooling all the i-vectors. The resulting vector of posterior probabilities for an i-vector $\vec{i}$ represent a condition quality vector (q-vector) $\vec{q}$ [Fe12], with entries:

$$q(j) = \frac{P(\vec{i}\,|\,\Lambda_j)}{\sum_{j=1}^{55} P(\vec{i}\,|\,\Lambda_j)}. \tag{1}$$

2.4 Neural Network schemes

Feed forward neural networks consist of layers of units [Bi06]. An input layer and an output layer are linked over a number of hidden layers by numerous connections, where the connections between units of each layer are weighted. In [He15], initial weights are proposed having a standard deviation of $\sqrt{2/n_l}$, with n_l being the number of incoming connections to the unit. In each unit, a linear combination, the *response*, is constructed from the outputs of the previous layer's units. A non-linear activation function is evaluated on the response to achieve the output, or *activation*, of the units e.g., the *linear rectifier, ReLU* activation function [LBH15] and the sigmoid function for bounded activations [Bi06]. Networks are trained to optimize the performance regarding the cost function using gradient descent, where the Adam algorithm [KB14] and *backpropagation* [Bi06] can be employed. As a cost function, the binary cross-entropy function is a measure of the distance between the distribution of the actual classes and the distribution of the prediction. In this work, we utilize a single-unit output layer, representing a system's score. In order to avoid over-fitting of the training data, different regularization schemes can be employed, such as *weight decay* [Bi06], *dropout* [Sr14], and *batch normalization* [IS15].

3 Deep Quality-informed Normalization

In order to account for cohort-related data as well as quality information, we propose to construct the input layer to a feed forward neural network based on the comparison

score, reference and probe i-vectors $\vec{i}_{ref}, \vec{i}_{prb}$ as well as corresponding q-vectors $\vec{q}_{ref}, \vec{q}_{prb}$, cf. Fig. 1, whereas a normalized score between 0 and 1, representing impostor and genuine classes, respectively, is obtained via a single unit output layer with a sigmoid activation function, yielding rather discriminative than well-calibrated scores. By training the network on the cohort data set, we assume the network model to comprise cohort and quality information, whilst achieving anonymity (not only pseudonymization) for the cohort speakers. Furthermore, massive data amounts featuring multi-condition quality is not required to be transferred to each mobile device by the biometric system operator.

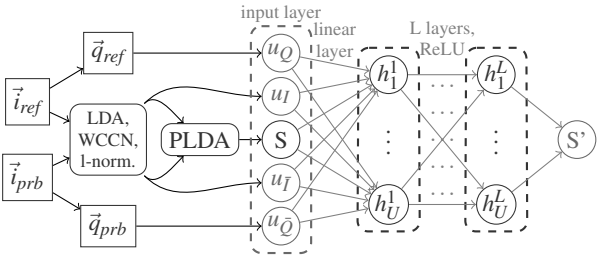


Fig. 1: Proposed deep quality-informed normalization network design with input layer (green), hidden layers (red), and output layer score S'. $u_Q, u_{\bar{Q}}$ represent Q-dim. q-vectors, and $u_I, u_{\bar{I}}$ represent I-dim. i-vectors, respectively.

For the purpose of accounting for linear normalization approaches e.g., linear quality calibration [Bd11, Fe12, Na16], the first hidden layer of the proposed network employs a linear activation function $f(x) = a + bx$. During training, input features are adaptively normalized with respect to the amount of genuine and impostor comparisons. Deeper hidden layers are non-linear using the ReLU activation function. The weights are initialized by the scheme proposed in [He15]. Convergence is reached after 3 epochs on a random-selected 20% held-out validation subset, on which the best performing model is chosen. In order to achieve an effective class balance of equal priors, genuine comparisons are weighted higher than the impostor comparisons during network training. The network configuration is referred to as (L,U) with a network of a linear layer with U units, followed by L non-linear layers of U units, cf. Fig. 1.

4 Experimental Set-Up and Analysis

For the purpose of studying the proposed method, first we examine regularization impacts on a fixed configuration of number of layers and units, finding $\lambda = 10^{-5}$ to reduce overfitting well, then parameters of the deep neural network with fix regularization parameter are examined comparing reasonable configurations on the testing set. In order to gain insights on the robustness of the proposed normalization scheme, a sensitivity analysis is conducted by excluding poor quality conditions from training the normalization network.

Implementations are based on Python 2.7 with Keras 1.1.1 and Theano 0.9.0.dev1, Matlab 2016b, and the BOSARIS toolkit [Bd11]. The data used is the same as in [Na15] of the I4U file list for NIST SRE'12 [Sa13]. The dataset consists of 55 different degradations in duration and noise type and level, denoted here as degradation conditions, cf.

Tab. 1. There are 680 reference i-vectors and 357269 probe i-vectors in the training dataset, and 723 reference and 388278 probe i-vectors in the test set. The i-vectors are processed dependent on the noise condition by performing linear discriminant analysis (LDA) to 200 dimensions, within class covariance normalization (WCCN), and length normalization [GREW11]. Baseline scores are derived from our recent work [Na15, Na16].

As an application-independent performance metric, we use minimum cost of log-likelihood ratio scores C_{llr}^{min} [BdP08], i.e. the generalized empirical cross-entropy of genuine and impostor scores, assuming well-calibrated systems in terms of Bayes decisions. The upper bound of C_{llr}^{min} is determined by the EER of the ROC's convex hull [Bd11].

4.1 Experimental Analysis: Network Configuration

In order to examine network configurations, we investigate on $L = 1, 2, 4$ layers, where all layers comprise the same amount of hidden units, i.e. $U = 50, 100, 200$ units. Tab. 2 compares the different networks on the test set: configuration (1, 50) yields the largest condition-average C_{llr}^{min} gain over a conventional i-vector / PLDA baseline system of 6.2% with the lowest standard variation, i.e. with rather stable improvements among all conditions. Configuration (2, 100) yields the second largest gains regarding average and deviation in terms of C_{llr}^{min}, but also regarding pooled-condition performance, where the (2, 50) network yields the largest gains. Accounting for potential over-fitting, dropout is examined on (1, 50) and (2, 100) networks with a 20% dropout rate: on average, C_{llr}^{min} grows, which may occur due to a too high dropout rate. Further investigations are carried out on the (1, 50) configuration, due its gains on pooled performance.

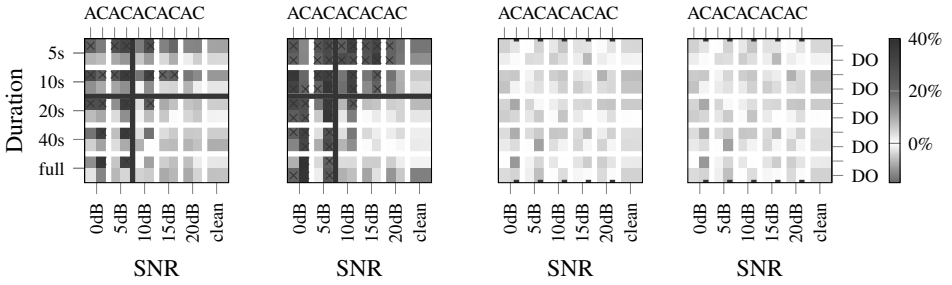
Tab. 2: Benchmark of relative C_{llr}^{min} changes (in %) to PLDA baseline on the test set regarding condition averaging (μ), standard deviation (σ), and pooling (p), and dropout training (DO).

(L, U)	(1, 50)	(1, 100)	(1, 200)	(2, 50)	(2, 100)	(2, 200)	(4, 50)	(4, 100)	(4, 200)	(1, 50) DO	(2, 100) DO
μ	**-6.2**	1.4	-2.0	-2.1	***-5.7***	0.8	-2.9	-5.2	-0.9	*1.3*	4.9
σ	**2.4**	6.6	3.5	4.3	***2.6***	4.4	3.1	2.9	3.8	*1.6*	4.0
p	-4.6	0.9	-0.2	**-6.6**	***-6.4***	0.4	-0.2	-3.4	0.0	*-2.5*	7.1

4.2 Robustness Analysis to unseen signal degradation and noise types

For the purpose of examining the robustness of the proposed normalization, training is conducted with unseen test conditions, i.e. all conditions afflicted with SNR levels $\leq 5\ dB$ and with durations $\leq 10\ s$ are excluded. Figs. 2a, 2b compare the effects to (1, 50) and (2, 100) configurations, with and without employing dropout, regarding whether or not the C_{llr}^{min} performance is not exceeding a $\pm 20\%$ performance band with respect to each condition's C_{llr}^{min}. In this analysis, the (1, 50) configuration outperforms the (2, 100) in terms

of coherence stability. Also, employing dropouts sustain coherent and stable performance. By placing focus on robustness towards noise type rather than low-SNRs, we exclude all CROWD noise afflicted conditions from training instead: both configurations perform stable and coherent with slight benefits from conducting dropout training, see Figs. 2c, 2d.



(a) conf. (1, 50), low-SNR and short duration conditions excluded

(b) conf. (2, 100), low-SNR and short duration conditions excluded

(c) conf. (1, 50), CROWD noise conditions excluded

(d) conf. (2, 100), CROWD noise conditions excluded

Fig. 2: Relative C_{llr}^{min} change on test set (in %). Performance by duration and SNR regarding AC (A) and CROWD (C) noise as well as whether dropout is conducted (DO). Green lines indicate the conditions excluded from training. Crosses denote relative C_{llr}^{min} changes above $\pm 20\%$.

4.3 Summary and Discussion

Examining deeper architectures considering non-linear layers, gains compared to the baseline PLDA performance are observed on average, though not further increasing the single linear layer performance. Comparatively, the cohort normalization in [Na15] yields up to 8.2% relative gains in C_{llr}^{min} on single conditions. In the robustness analysis, i.e. by excluding poor quality conditions and the more challenging noise type, the proposed approach reveals to benefit on good quality conditions, the performance of the (1, 50) configuration is preserved within a $\pm 20\%$ performance band on unseen poor quality conditions. Contrastively, on excluding overlapping speech (CROWD noise) conditions, either (1, 50) and (2, 100) configurations perform comparatively stable. Thus, the proposed approach benefits rather from training on a broad scale of SNR levels than on more noise types, posing a challenging scenario due to overlapping biometric features of other subjects.

5 Conclusion

In this study, we introduced a neural network based normalization approach utilizing quality estimates, suitable for unconstrained environments under data privacy as well as limited resource concerns regarding the data of cohort speakers. As system operators transmit trained networks to mobile devices instead of cohort data, data privacy is achieved for cohort subjects, while sustaining comparative discrimination performance. Robustness analyses show benefits of knowing levels of SNR levels and durations during training over knowing different noise types of mid / high-SNR levls during training.

6 Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research and the Arts within the Center for Research in Security and Privacy (`www.crisp-da.de`), and the BioMobile II project (no. 518/16-30).

References

[BBM13] Bousquet, P.-M.; Bonastre, J.-F.; Matrouf, D.: Identify the Benefits of the Different Steps in an i-Vector Based Speaker. Springer-Verlag Berlin Heidelberg, chapter CIARP, Part II, pp. 278–285, 2013.

[Bd11] Brümmer, N.; de Villiers, E.: The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing. Technical report, AGNITIO Research, South Africa, December 2011. Last accessed: 2017-05-15.

[BdP08] Brümmer, N.; du Preez, J.: Application-Independent Evaluation of Speaker Detection. Computer Speech and Language, 20(2):230–275, July 2008.

[Bi06] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, 2006.

[Br16] Brümmer, N.; Swart, A.; Jorrím-Prieto, J.; García, P. et al.: ABC NIST SRE 2016 System Description. In: Proc. of the NIST SRE 2016 workshop. 2016.

[CL14] Cumani, S.; Laface, P.: Generative pairwise models for Speaker Recognition. In: Proc. Odyssey 2014: The Speaker and Language Recognition Workshop. 2014.

[Cu11] Cumani, S.; Batzu, P. Domenico; Colibro, D.; Vair, C.; Laface, P.; Vasilakakis, V.: Comparison of Speaker Recognition Approaches for Real Applications. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). 2011.

[De11] Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transaction on Audio, Speech, and Language Processing (TASLP), 19(4):788–798, May 2011.

[Eu16] European Council: , Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), April 2016.

[Fe12] Ferrer, L.; Burget, L.; Plchot, O.; Scheffer, N.: A Unified Approach for Audio Characterization and its Application to Speaker Recognition. In: Odyssey 2012: The Speaker and Language Recognition Workshop. 2012.

[GREW11] Garcia-Romero, D.; Epsy-Wilson, C.Y.: Analysis of i-vector length normalization in Speaker Recognition systems. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). pp. 249–252, 2011.

[Ha13] Hasan, T.; Saeidi, R.; Hansen, J. H. L.; van Leeuwen, D. A.: Duration Mismatch Compensation for i-vector based Speaker Recognition systems. In: Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2013.

[He15] He, K.; Zhang, X.; Ren, S.; Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE Intl. Conf. on Computer Vision. pp. 1026–1034, 2015.

[IS15] Ioffe, S.; Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[KB14] Kingma, D.; Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[Ke05] Kenny, P.: Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Technical Report CRIM-06/08-13, CRIM, Montreal, 2005.

[LBH15] LeCun, Y.; Bengio, Y.; Hinton, G.: Deep learning. Nature, May 2015.

[Ma13] Mandasari, M. I.; Saeidi, R.; McLaren, M.; van Leeuwen, D. A.: Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions. IEEE Trans. on Audio, Speech and Language Processing (TASLP), 21(11):2425–2438, 2013.

[MSvL15] Mandasari, M. I.; Saeidi, R.; van Leeuwen, D. A.: Quality measures based Calibration with Duration and noise dependency for Speaker Recognition. Speech Communication, 72:126–137, September 2015.

[Na15] Nautsch, A.; Saeidi, R.; Rathgeb, C.; Busch, C.: Analysis of mutual Duration and noise effects in Speaker Recognition: benefits of condition-matched cohort selection in Score Normalization. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). pp. 3006–3010, 2015.

[Na16] Nautsch, A.; Saeidi, R.; Rathgeb, C.; Busch, C.: Robustness of Quality-based Score Calibration of Speaker Recognition Systems with respect to low-SNR and short-Duration conditions. In: Proc. of Odyssey 2016: The Speaker and Language Recognition Workshop. pp. 358–365, 2016.

[RQD00] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Conversational Speech, Digital Signal Processing, 10:19–41, 2000.

[Sa13] Saeidi, R.; Lee, K.A.; Kinnunen, T. et al.: I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification. In: Proc. of the Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH). ISCA, 2013.

[SR05] Sturim, D.E.; Reynolds, D.A.: Speaker adaptive Cohort Selection for tnorm in text-dependent Speaker Verification. In: Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). 2005.

[Sr14] Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1):1929–1958, 2014.

[Va16] Vair, C.; Colibro, D.; Dalmasso, E.; Farrell, K. et al.: Nuance - Politecnico di Torino (NPT) System Description for NIST 2016 Speaker Recognition Evaluation. In: Proc. of the NIST SRE 2016 workshop. 2016.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 251

Evaluation of CNN architectures for gait recognition based on optical flow maps

F.M. Castro[1], M.J. Marín-Jiménez[2], N. Guil[1], S. López-Tapia[3], N. Pérez de la Blanca [3]

Abstract: This work targets people identification in video based on the way they walk (*i.e.*gait) by using deep learning architectures. We explore the use of convolutional neural networks (CNN) for learning high-level descriptors from low-level motion features (*i.e.*optical flow components). The low number of training samples for each subject and the use of a test set containing subjects different from the training ones makes the search of a good CNN architecture a challenging task. We carry out a thorough experimental evaluation deploying and analyzing four distinct CNN models with different depth but similar complexity. We show that even the simplest CNN models greatly improve the results using shallow classifiers. All our experiments have been carried out on the challenging TUM-GAID dataset, which contains people in different covariate scenarios (*i.e.*clothing, shoes, bags).

Keywords: Deep Neural Networks, Gait Recognition, Optical Flow, ResNet, 3D-CNN.

1 Introduction

The goal of *gait recognition* is to identify people by the way they walk. This type of biometric approach is considered non-invasive, since it is performed at a distance, and does not require the cooperation of the subject that has to be identified, in contrast to other methods as iris- or fingerprint-based approaches. Gait recognition has application in the context of video surveillance, ranging from control access in restricted areas to early detection of persons of interest as, for example, v.i.p. customers in a bank office.

In last years, great effort has been put into the problem of people identification based on gait patterns [Hu04]. However, previous approaches have mostly used hand-crafted features for representing the human gait [BD09, HB06, Ca17], which do not easily adapt to diverse datasets, due to the specificity of the hand-crafted descriptors obtained for each dataset. Therefore, we propose an end-to-end approach based on convolutional neural networks that given low-level optical flow maps, directly extracted from video frames (see Fig. 1), is able to learn and extract higher-level features suitable for representing human gait: *gait signature*. In addition, we also present a fair comparative between four models based on three of the most popular kinds of CNN architectures used in computer vision tasks: LeNet [LB95], VGG [SZ14] and ResNet [He16]. The contribution of this paper is twofold: (*i*) a set of CNN models for gait recognition using optical flow; and, (*ii*) a thorough experimental study to validate the proposed models on the standard TUM-GAID dataset for gait identification, obtaining state-of-the-art results.

The rest of the paper is organized as follows. We continue by reviewing the related work. Then, Sec. 2 explains our four different models for learning gait signatures and identifying

[1] University of Málaga, Department of Computer Architecture, Spain
[2] University of Córdoba, Department of Computing and Numerical Analysis, Spain
[3] University of Granada, Department of Computer Science and Artificial Intelligence, Spain

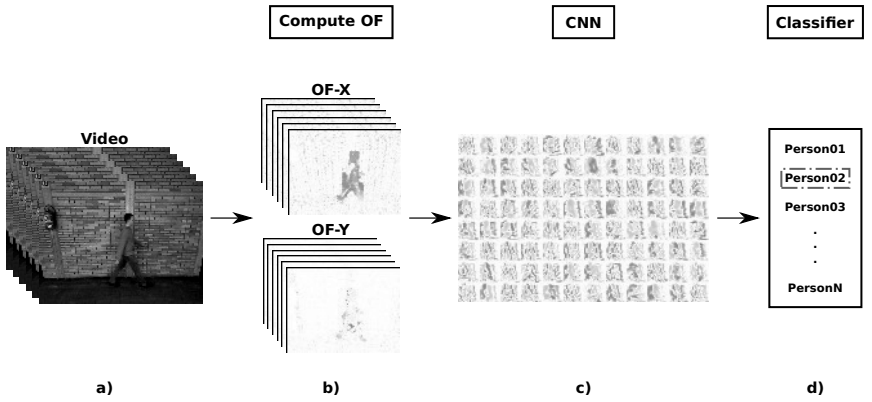


Fig. 1: **Pipeline for gait recognition**. a) The input is a sequence of RGB video frames. b) Optical flow is computed along the sequence. c) Optical flow subsequences are passed through the CNN to obtain gait signatures. e) Classification of the extracted gait signatures. Note: positive flows are in pink and negative flows in blue. (Best viewed in colour).

people. Sec. 3 contains the experiments and results. Finally, we present the conclusions and future work in Sec. 4.

1.1 Related work

Traditionally, deep learning approaches based on Convolutional Neural Networks (CNN) have been used in image-based tasks with great success [KSH12]. In the last years, deep architectures for video have appeared, specially focused on action recognition, where the inputs of the CNN are subsequences of stacked frames. In [SZ14], Simonyan and Zisserman proposed to use as input to a CNN a volume obtained as the concatenation of frames with two channels that contain the optical flow in the x-axis and y-axis respectively. To normalize the size of the inputs, they split the original sequences into subsequences of 10 frames, considering each subsample independently. A natural modification is presented by Ji *et al.* [Ji13], where a 3D convolutional network is developed to capture temporal information from multiple frames. Then, Tran *et al.* [Tr15] propose a new 3D network which uses raw videos as input, instead of preprocessed inputs. Recently, a new approach has been developed by He *et al.* [He16]. They propose a new kind of CNN which has a large number of layers and residual connections to avoid the vanishing gradient problem. Although several papers can be found for the task of human action recognition using deep learning techniques, it is hard to find such type of approaches applied to the problem of gait recognition. In [HC13], Hossain and Chetty propose the use of Restricted Boltzmann Machines to extract gait features from binary silhouettes, but a very small probe set (*i.e.*only ten different subjects) was used for validating their approach. A more recent work, [WHW15], uses a random set of binary silhouettes from a sequence to train a CNN that accumulates the calculated features to achieve a global representation of the dataset. In [AM15], raw 2D GEI are employed to train a simple CNN for gait recognition. A more complex work is presented in [GB15] where GEI are used to train an ensemble of CNN and

This work has been funded under projects TIC-1692 (Junta de Andalucía) and TIN2016-75279-P (Spanish Ministry of Science and Tech.). The GPU Titan X Pascal used for this research was donated by NVIDIA.

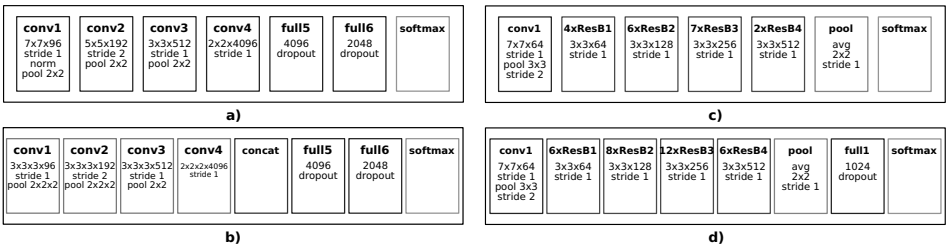


Fig. 2: **Proposed CNN models for gait signature extraction**. **a) 2D-CNN:** linear CNN with four 2D convolutions, two fully connected layers and a softmax classifier. **b) 3D-CNN:** four 3D convolutions, two fully connected layers and a softmax classifier. **c) ResNet-A:** residual CNN with a 2D convolution, four residual blocks, an average pooling layer and a final softmax classifier. **d) ResNet-B:** extended version of ResNet-A. Note that before the first block of each kind (ResB 1, 2, 3, 4), there is an adapter convolution to resize the input image to the size of the next block.

a Multilayer Perceptron is employed as classifier. In [Wu17], given two GEI descriptors, they learn a metric to decide whether both descriptors belong to the same subject or not. All those previous CNN-based approaches propose precomputed GEI descriptors as input features. In contrast, our approach builds a spatio-temporal volume of optical flow [SZ14] as input to a CNN specially designed for gait recognition, what will allow the CNN to learn characteristic gait patterns directly from the source, *i.e.*the motion.

2 Proposed approach

In this section we describe our proposed framework to address the problem of gait recognition using CNN. The proposed pipeline is represented in Fig. 1: *(i)* compute optical flow (OF) along the whole sequence; *(ii)* build up a data cuboid from consecutive OF maps; *(iii)* feed the different CNNs with an OF cuboid to extract the gait signature; and, *(iv)* using the gait signature, decide the subject identity.

2.1 Input data

The use of optical flow (OF) as input data for action representation in video with CNN has already shown excellent results [SZ14]. Nevertheless human action is represented by a wide, and usually well defined, set of local motions. In our case, the set of motions differentiating one gait style from another is much more subtle and local.

Let F_t be an OF map computed at time t and, therefore, $F_t(x,y,c)$ be the value of the OF vector component c located at coordinates (x,y), where c can be either the horizontal or vertical component of the corresponding OF vector. The input data I_L for the CNN are cuboids built by stacking L consecutive OF maps F_t, where $I_L(x,y,2k-1)$ and $I_L(x,y,2k)$ corresponds to the value of the horizontal and vertical OF components located at spatial position (x,y) and time k, respectively, ranging k in the interval $[1,L]$.

Since original video sequences have different temporal length, and CNN requires a fixed size input, we extract subsequences of L frames from the full-length sequences.

2.2 CNN architectures for gait signature extraction

We have selected three of the architectures that most frequently appear in the bibliography and produce state-of-the-art results in different topics (*e.g.*action recognition, object detection, etc). The proposed architectures are: *(i)* the LeNet architecture [LB95], adapted to a

model named (*2D-CNN*), which is the most common architecture; *(ii)* the VGG architecture [SZ14], adapted to use 3D convolutions on optical flow inputs and named (*3D-CNN*), which is specially designed to capture information in video sequences; and, *(iii)* two CNN models with residual units (named *ResNet* [He16]), used to experiment with deeper models on this task, as the network depth has been recently pointed out as one the most relevant factors to achieve the state of the art in many tasks [KSH12].

To carry out a fair comparison, three of the four models have been designed to have a similar number of parameters, where the 2D-CNN model has been taken as a reference (*i.e.*$\sim 18.5M$). This choice allows us to carry out a comparative study which is independent of the network capacity. Due to the particular design of the fourth one, it has a different number of parameters.

We describe below the four models compared in the experimental section (Sec. 3):

2D-CNN (16 layers): This CNN is composed of the sequence of layers shown in Fig. 2.a). All convolutional layers use a ReLU function and all *conv* blocks contain a max-pooling operation.

3D-CNN (16 layers): As optical flow has two components and the CNN uses temporal kernels, the network is split into two branches: x-flow and y-flow. Therefore, each branch contains half of the total filters. Then, this CNN is composed of the sequence of layers shown in Fig. 2.b). Note that 'concat' layer concatenates both branches (x-flow and y-flow) into a single one. All convolutional layers use a ReLU function and all *conv* blocks contain max-pooling.

ResNet-A (167 layers): This CNN is composed of the sequence of layers and residual blocks (a sequence of two convolutions of size 3×3 and a sum layer, as defined in [He16]) shown in Fig. 2.c). As our model follows the indications defined in [He16], we only describe the main blocks. Note that all convolutional layers use the rectification (ReLU) activation function and batch normalization.

ResNet-B (268 layers): This CNN is an extended version of ResNet-A, composed of the sequence of layers and residual blocks shown in Fig. 2.d). Note that all convolutional layers use the parametric rectification (PReLU) [He15] activation function, local response normalization (LRN) and batch normalization. The use of PReLU is specially useful in our case as optical flow has negative components which contain important information about motion. Therefore, the network uses more information and the gradients are more powerful, avoiding the vanishing gradient problem.

2.3 Training details

For models 2D-CNN, 3D-CNN and ResNet-A, during training, the weights are learnt using mini-batch stochastic descent algorithm with momentum equal to 0.9. We set weight decay to $5 \cdot 10^{-4}$ and dropout to 0.4 (2D-CNN and 3D-CNN). The learning rate is initially set to 10^{-2} and divided by 10 when the validation error gets stuck. At each epoch, a mini-batch of 150 samples is constructed by random selection over a balanced training set (*i.e.*almost same proportion of samples per class).

As ResNet-B has some peculiarities, training parameters must be adapted. In this case, mini batches of size 64 are used. The learning rate policy follows a triangular scheme that

consists of varying the learning rate between a minimum and a maximum value following a triangular pattern with the training iterations. The triangular learning rate parameters range from 0.003 to 0.015 during 4 epochs. The model was trained with a total of 24 epochs. Finally, dropout is used before each fully connected layer with a value of 0.1. Also weight decay regularization with value 0.0005 was imposed. Note that all hyperparameters have been cross-validated and only the best ones are presented in this paper.

3 Experiments and results

3.1 Dataset

TUM-GAID [Ho14] contains 305 subjects walking on four different conditions: normal walking (*N*), carrying a backpack (*B*), wearing coating shoes (*S*) and elapsed time (*TN*, *TB*, *TS*). We follow the standard experimental protocol defined by the authors of the dataset [Ho14]. Therefore, we use 100 subjects as training set, 50 different subjects as validation set and 155 different subjects as test set – note that it is distinguished between 'subject partitions' and 'sequence partitions', *i.e.*, for each subject, training, validation and test sequences are available. As we have different subjects between training and testing, it is needed to fine-tune the model with four training sequences of normal walking of the test subject partition. Note that the sequences used for fine-tuning are not used during testing. For testing, we use six sequences that have never been seen before by our model according to the partitions defined in [Ho14].

3.2 Implementation details

All videos are resized to a common resolution of 80×60 pixels, keeping the original aspect ratio of the video frames. Given the resized video sequences, we compute dense OF on pairs of frames by using the method of Farneback [Fa03] implemented in OpenCV library. In parallel, people are located in a rough manner along the video sequences by background subtraction [KB02]. Then, we crop the video frames to remove part of the background, obtaining video frames of 60×60 pixels (full height is kept) and to align the subsequences (people are x-located in the middle of the central frame). Finally, from the cropped OF maps, we build subsequences of 25 frames by stacking OF maps with an overlap of $\Theta\%$ frames. As this dataset is relatively small, we need to choose an intermediate overlapping rate value that allows to obtain training samples with enough variability between them. In our case, we empirically choose $\Theta = 80\%$, that is, to build a new subsequence, we use 20 frames of the previous subsequence and 5 new frames. For most state-of-the-start datasets, 25 frames cover almost one complete gait cycle, as stated by other authors [BD09]. Therefore, each OF volume has size $60 \times 60 \times 50$.

To increase the amount of training samples we add mirror sequences and apply spatial displacements of ± 5 pixels per axis, obtaining a total of 8 new samples from each original one. Then, mirror sequences are computed, obtaining about $270k$ training samples. Note that in Sec. 2.1, we split the whole video sequence into overlapping subsequences of a fixed length, and those subsequences are classified independently. Therefore, in order to derive a final identity for the whole sequence, we multiply the probabilities returned by the Softmax layer for all subsequences of the same sequence. Before feeding each sample into the CNN, the mean value of the whole training dataset is subtracted.

We ran our experiments on a PC with 32 cores at 2.2 GHz, 256 GB of RAM and a GPU NVIDIA Titan X Pascal, with MatConvNet library [VL15] running on Matlab 2016a for Ubuntu 14.04 and Caffe [Ji14] library for ResNet-B.

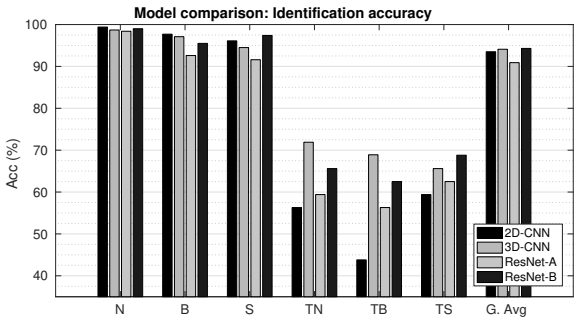


Fig. 3: **Model comparison in terms of identification accuracy**. Results grouped per scenario: normal 'N', backpacks 'B', shoes 'S' and temporal cases 'T*x*'. Group 'G.Avg' corresponds to global average on the six scenarios.

3.3 Experimental results

After splitting the training sequences (of the training subjects) into subsequences, we got a training set composed of 269352 samples used for learning the filters; and a second training set composed of 108522 samples for training the softmax layer from the subset of test subjects. Test sequences are never used for training or validation of the model.

Fig. 3 offers a visual comparison of the results obtained with each of the four tested architectures grouped per scenario type. In terms of scenario type, note that the temporal ones (T*x*) are the most challenging, as there exists a large change in subject appearance with regard to the non-temporal cases where the filters of the networks were trained.

To put our results in context, Tab. 1 contains the state-of-the-art and the comparison between the four different models (rows '2D-CNN', '3D-CNN', 'ResNet-A' and 'ResNet-B'). We have applied the PFM descriptor [Ca17] on resized videos of 80×60 to obtain a fair comparison. Comparing the CNN results with the state-of-the-art , 2D-CNN achieves on average the best results for the *non-temporal* scenarios. For the *temporal* cases, 3D-CNN obtains the best results. On global average (column '*G.Avg*'), ResNet-B sets a new state-of-the-art with an accuracy 0.2% better than the rest of CNNs and 6.1% better than the best handcrafted method. Note that CNNs use an input 16 times lower than the rest of the compared methods.

4 Discussion and Conclusions

The relevance of the complexity in CNN architectures, when applied to the gait recognition task, has been analysed through a comparative study of four models (from three deep architectures) and its comparison to results from methods based on handcrafted features. The first conclusion is that in this task, as in many others, the deep CNN architectures overcome shallow and handcrafted methods. This fact points out the importance of the architecture depth to extract relevant features. The second conclusion is that the four deep models achieve similar results in the *non-temporal* scenario, but in the *temporal* one the differences are more significant. The filters used by the 3D-CNN model make the difference in the *temporal* scenario. The standard convolutional architectures obtain the best results on the *non-temporal* and *temporal* scenarios as its design is focused on the main

	Method	*N*	*B*	*S*	*Avg*	*TN*	*TB*	*TS*	*Avg*	*G. Avg*
640 × 480	GEI [Ho14]	99.4	27.1	52.6	59.7	44.0	6.0	9.0	19.7	56.0
	SEIM [WBR14]	99.0	18.4	96.1	71.2	15.6	3.1	28.1	15.6	66.6
	GVI [WBR14]	99.0	47.7	94.5	80.4	62.5	15.6	62.5	46.9	77.3
	SVIM [WBR14]	98.4	64.2	91.6	84.7	65.6	31.3	50.0	49.0	81.4
	RSM [GL13]	100	79.0	97.0	92.0	58.0	38.0	57.0	51.3	88.2
	PFM [Ca17]	75.8	70.3	32.3	59.5	50.0	40.6	25.0	38.5	57.5
80 × 60	2D-CNN	99.4	97.7	96.1	**97.7**	56.3	43.8	59.4	53.2	93.5
	3D-CNN	98.7	97.1	94.5	96.7	71.9	68.9	65.6	**68.8**	94.1
	ResNet-A	98.4	92.6	91.6	94.2	59.4	56.3	62.5	59.4	90.9
	ResNet-B	99.0	95.5	97.4	97.3	65.6	62.5	68.8	65.6	**94.3**

Tab. 1: **State-of-the-art on TUM GAID**. Percentage of correct recognition on TUM-GAID for diverse methods published in the literature. Bottom rows correspond to our proposal, where instead of using video frames at 640×480, a resolution of 80×60 is used. Each column corresponds to either a different scenario or average on scenarios (*i.e.Avg*, *G.Avg*). Best results are marked in bold.

variations of the signal, spatial in 2D-CNN and temporal in 3D-CNN. Regarding the two ResNet models there are many differences between them in terms of design (see Fig.2) and training parameters. The ResNet-B model is a much more deeper architecture needing of PReLU activations and adaptive learning rate to obtain a good optimum. A final fully connected layer with dropout was added as well. Nevertheless and despite all these improvements, an increment of only 3.4 points in score is obtained w.r.t. ResNet-A. This result shows that the addition of residual layers although allows to fit deeper models, needs of a good learning rate policy to obtain a good optimum. The ResNet architecture achieves the overall best results when it is properly fitted. Our results reinforce, for the gait recognition task, the empirical finding of other works that indicates that architectures with enough depth are needed in order to obtain high classification accuracy. In addition, the use of appropriate activation functions has also shown to be a very relevant choice on this task. Focusing on the training speed, independently of the number of parameters, 3D-CNN needs more training time, followed by 2D-CNN and ResNet which is the fastest one.

As future work, we plan to extend our study to identify the kind of architectures more suitable to combine motion with appearance (*i.e.*RGB data), applying them to more gait datasets in which optical flow can be computed – this would allow us to perform transfer learning between networks trained on different data.

References

[AM15] Alotaibi, M.; Mahmood, A.: Improved Gait recognition based on specialized deep convolutional neural networks. In: AIPR Workshop. pp. 1–7, 2015.

[BD09] Barnich, Olivier; Droogenbroeck, Marc Van: Frontal-view gait recognition by intra- and inter-frame rectangle size distribution. Patt. Recogn. Letters, 30(10):893 – 901, 2009.

[Ca17] Castro, Francisco M.; Marín-Jiménez, M.J.; Guil Mata, N.; Muñoz Salinas, R.: Fisher Motion Descriptor for Multiview Gait Recognition. Intl. J. of Patt. Recogn. in Artificial Intelligence, 31(1), 2017.

[Fa03] Farnebäck, Gunnar: Two-Frame Motion Estimation Based on Polynomial Expansion. In: Proc. of Scandinavian Conf. on Image Analysis. volume 2749, pp. 363–370, 2003.

[GB15] Gálai, Bence; Benedek, Csaba: Feature selection for Lidar-based gait recognition. In: Computational Intelligence for Multimedia Understanding (IWCIM). pp. 1–5, 2015.

[GL13] Guan, Yu; Li, Chang-Tsun: A robust speed-invariant gait recognition system for walker and runner identification. In: Intl. Conf. on Biometrics (ICB). pp. 1–8, 2013.

[HB06] Han, Ju; Bhanu, Bir: Individual recognition using gait energy image. IEEE PAMI, 28(2):316–322, 2006.

[HC13] Hossain, Emdad; Chetty, Girija: Multimodal Feature Learning for Gait Biometric Based Human Identity Recognition. In: NIPS. pp. 721–728, 2013.

[He15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: ICCV. pp. 1026–1034, 2015.

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

[Ho14] Hofmann, Martin; Geiger, Jrgen; Bachmann, Sebastian; Schuller, Bjrn; Rigoll, Gerhard: The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. J. of Visual Com. and Image Repres., 25(1):195 – 206, 2014.

[Hu04] Hu, Weiming; Tan, Tieniu; Wang, Liang; Maybank, Steve: A survey on visual surveillance of object motion and behaviors. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 34(3):334–352, 2004.

[Ji13] Ji, S.; Xu, W.; Yang, M.; Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition. IEEE PAMI, 35(1):221–231, Jan 2013.

[Ji14] Jia, Yangqing; Shelhamer, Evan; Donahue, Jeff; Karayev, Sergey; Long, Jonathan; Girshick, Ross; Guadarrama, Sergio; Darrell, Trevor: Caffe: Convolutional Architecture for Fast Feature Embedding. arXiv preprint arXiv:1408.5093, 2014.

[KB02] KaewTraKulPong, P.; Bowden, R.: An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection. In: Video-Based Surveillance Systems, pp. 135–144. 2002.

[KSH12] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E.: ImageNet Classification with Deep Convolutional Neural Networks. In: NIPS. pp. 1097–1105, 2012.

[LB95] LeCun, Yann; Bengio, Yoshua: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

[SZ14] Simonyan, Karen; Zisserman, Andrew: Two-stream convolutional networks for action recognition in videos. In: NIPS. pp. 568–576, 2014.

[Tr15] Tran, Du; Bourdev, Lubomir D.; Fergus, Rob; Torresani, Lorenzo; Paluri, Manohar: Learning Spatiotemporal Features with 3D Convolutional Networks. In: ICCV. IEEE, 2015.

[VL15] Vedaldi, A.; Lenc, K.: MatConvNet – Convolutional Neural Networks for MATLAB. In: Proceeding of the ACM Int. Conf. on Multimedia. 2015.

[WBR14] Whytock, Tenika; Belyaev, Alexander; Robertson, NeilM.: Dynamic Distance-Based Shape Features for Gait Recognition. Journal of Mathematical Imaging and Vision, 50(3):314–326, 2014.

[WHW15] Wu, Zifeng; Huang, Yongzhen; Wang, Liang: Learning Representative Deep Features for Image Set Analysis. IEEE Trans. on Multimedia, 17(11):1960–1968, Nov 2015.

[Wu17] Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T.: A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. IEEE PAMI, 39(2):209–226, 2017.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 259

How Random is a Classifier given its Area under Curve?

Chris Zeinstra[1], Raymond Veldhuis[1], Luuk Spreeuwers[1]

Abstract: When the performance of a classifier is empirically evaluated, the Area Under Curve (AUC) is commonly used as a one dimensional performance measure. In general, the focus is on good performance (AUC towards 1). In this paper, we study the other side of the performance spectrum (AUC towards 0.50) as we are interested to which extend a classifier is random given its AUC. We present the *exact* probability distribution of the AUC of a truely random classifier, given a finite number of *distinct* genuine and imposter scores. It quantifies the "randomness" of the measured AUC. The distribution involves the restricted partition function, a well studied function in number theory. Although other work exists that considers confidence bounds on the AUC, the novelty is that we do not assume any underlying parametric or non-parametric model or specify an error rate. Also, in cases in which a limited number of scores is available, for example in forensic case work, the exact distribution can deviate from these models. For completeness, we also present an approximation using a normal distribution and confidence bounds on the AUC.

Keywords: Random Classifier, AUC, Exact Distribution, Approximation.

1 Introduction

The trade off between the False Match Rate (FMR) and True Match Rate (TMR) of a classifier while varying the decision threshold is commonly reported in a receiver operating characteristic (ROC) curve [Fa06]. There exist several one dimensional classifier performance measures that can be derived from its ROC curve, for example, the Equal Error Rate and the Area under Curve [HM82]. In this study, we consider the Area Under Curve (AUC) measure. An ideal classifier has AUC=1, whereas a random classifier has AUC=0.50. The AUC is equal to the probability that a randomly chosen genuine score is larger than a randomly chosen imposter score [HT01]. Also, the AUC can be interpreted as the Wilcoxon-Mann-Whitney statistic [MW47] when ordering the genuine and imposter scores produced by the classifier [HM82], [MG02].

In any empirical performance evaluation, only a finite number of genuine and imposter scores is available. Under the assumption that genuine and imposter scores are drawn from unknown probability densities, ultimately the AUC is also a random variable, having a probability distribution on its own. If we could replicate the experiment having the exact same number of genuine and imposter scores, we most likely would have obtained a different ROC curve and AUC. In particular, this implies that the performance evaluation might yield an AUC value that is not identified as being produced by a random classifier. This could occur in the case of a subject anchored approach to evidence evaluation in which the available number of scores is limited, see [Me06] for a general framework.

[1] University of Twente, Faculty of EEMCS, SCS Group, P.O.Box 217, 7500 AE Enschede, The Netherlands, {c.g.zeinstra,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

The probability distribution of the AUC of a random classifier is easily derived for trivial cases. More precisely, we assume that (a) this classifier draws genuine and imposter scores randomly from the *same* probability distribution and (b) the drawn scores are *distinct*. The last condition is a necessary technicality; if we for example assume that scores come from a continuous interval, this condition is typically met. Suppose we construct a ROC curve based on 1 genuine score g and n imposter scores i_k, $k = 1,\ldots,n$. We have $n+1$ possible orderings of the scores:

$$g < i_1 < \ldots < i_{n-1} < i_n \text{ to } i_1 < i_2 < \ldots < i_n < g. \tag{1}$$

Since g and i_k come from the same distribution, each sequence in (1) has equal probability $\frac{1}{n+1}$. If l ($l = 1,\ldots,n+1$) is the position of g in any sequence in (1), then its AUC is equal to $\frac{l-1}{n}$. Hence, each possible AUC has equal probability. The one-to-one mapping in this trivial 1 genuine/n imposter case between sequences and the AUC does not hold in general. For example, both i_1, g_1, g_2, i_2 and g_1, i_1, i_2, g_2 yield AUC=0.50, and the situation becomes rapidly complex when m and n attain values found in practice.

The contribution of this paper is the exact probability distribution of the AUC of the random classifier for *any* finite number of genuine and imposter scores. Also, we present an approximation. This work can be used in the situation when we want to determine the probability that a random classifier produces the measured AUC; this is of interest when the measured AUC is low or the total number of scores is limited.

The remainder of this article is structured as follows. In Section 2, we present related work. Since the general approach involves the restricted partition function, we present its definition in Section 3. In Section 4, we present two theorems regarding respectively the probability distribution of the AUC and an approximation. Section 5 presents some examples of the exact and an application of the approximation. In Section 6, we discuss the two theorems. Finally, in Section 7 we present our conclusion.

2 Related Work

As indicated before, this work fits in a larger framework that studies whether two AUC's are significantly different by constructing confidence intervals. This is not only of importance in decision theory, but also for clinical medicine and psychology studies in which treatments are compared. We present some of these studies here.

For example, the work of [CM04] analytically derives exact and estimated confidence intervals based on a statistical and combinatorical analysis, using a fixed error rate and the number of genuine and imposter scores. Our work only uses the number of genuine and imposter scores, assuming that they are drawn from the same probability distribution. Another approach is the use of parametric models to construct confidence intervals. For example score distributions have been modeled as normal [HSZ09], binormal [MHS98], exponential [To77], and Gamma [PA95], from which expressions for the confidence intervals can be derived. Their main issue is the influence of the parametric assumption on the

estimation of confidence intervals. To cater for that situation, several non-parametric methods have been explored, including Wilcoxon-Mann-Whitney and De-Long non-parametric interval [DDCP88]. The work of [QH08] compares nine non-parametric approaches in different simulation scenarios (moderate to good AUC and different combinations of genuine and imposter scores). They found that their own empirical likelihood approach [QZ06] has a good coverage in different scenarios. Several studies have shown that methods can be negatively influenced by the number of considered scores. For example, [OL98] found that asymptotic methods are less accurate in this situation; the study of [Ha10] shows how estimates for the AUC can differ significantly from the true value.

In summary, these studies emphasise on one hand the restriction of our work (random classifier) and on the other hand its uniqueness (exact distribution, depending on the number of genuine and imposter scores).

3 Partition functions

The partition function is an essential function in number theory, a branch of mathematics that studies properties of integers [An98]. A partition of a positive integer k is a decomposition of k as a sum of positive integers. The partition function p counts the number of different partitions of a positive integer, disregarding any permutations in the order of the terms. For example $p(5) = 7$, since

$$5 = 5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1. \quad (2)$$

It is customary to order the terms in a partition from the largest to the lowest value. This can be written more formally as $k_1 + \ldots + k_r = k,$ and $k_1 \geq k_2 \geq \cdots k_r$. Also, by convention, the domain of p is extended by including $p(0) = 1$ and $p(k) = 0$ for $k < 0$.

There exist different "restricted" versions of the partition function. In particular, one can limit the number and value of the terms of a partition. Let $p(n,m;k)$ be the number of partitions of k which have at most m terms, each having maximum value n. In the sequel, we refer to this function as "the" restricted partition function. For example, $p(4,2;5) = 2$, since the maximum value is 4 and the maximum number of terms is 2:

$$5 = 4+1 = 3+2. \quad (3)$$

The restricted partition function has a generating function:

$$\sum_{k=0}^{nm} p(n,m;k)q^k = \binom{m+n}{m}_q, \quad (4)$$

in which

$$\binom{m+n}{m}_q = \frac{\prod_{j=1}^{m+n}(1-q^j)}{\prod_{j=1}^{m}(1-q^j)\prod_{j=1}^{n}(1-q^j)} \quad (5)$$

is the Gaussian binomial coefficient [An74]. It generalises the binomial coefficient as for $\lim_{q\nearrow 1}$, (5) reverts to the standard binomial coefficient $\binom{k+l}{l}$. As an example, we expand

$p(4,2;k)$ for $k = 0, \cdots, 8$:

$$\sum_{k=0}^{8} p(n,m;k)q^k = \binom{6}{2}_q = \frac{\prod_{j=1}^{6}(1-q^j)}{\prod_{j=1}^{2}(1-q^j)\prod_{j=1}^{4}(1-q^j)} = \frac{(1-q^5)(1-q^6)}{(1-q)(1-q^2)}. \quad (6)$$

It is straightforward to verify that (6) is equal to $1+q+2q^2+2q^3+3q^4+2q^5+2q^6+q^7+q^8$. In particular, we observe that $p(4,2;5) = 2$ (the factor of q^5), a result that was also demonstrated by (3).

4 Exact and Approximative Distribution

We have the following theorem on the distribution of AUC.

Theorem 1. Given m genuine and n imposter scores, all distinct, the possible values for AUC are

$$\{\frac{k}{mn} | k \in \{0,\ldots,mn\}\}. \quad (7)$$

Moreover, if the genuine and imposter scores are drawn from the *same* score distribution, then the probability distribution of the AUC is given by

$$\mathrm{p}\left(\mathrm{AUC} = \frac{k}{mn}\right) = \frac{p(n,m;k)}{\binom{n+m}{n}}, \quad (8)$$

where $p(n,m;k)$ is the restricted version of the partition function.

Proof. Having m genuine and n imposter scores, this divides the TMR (resp. FMR) space into $m+1$ (resp. $n+1$) points with distance $\frac{1}{m}$ (resp. $\frac{1}{n}$). Since we have distinct scores, whenever the threshold increases and passes a score, the corresponding operating point in ROC space will either move to the left with a step size $\frac{1}{n}$ or down with a step size $\frac{1}{m}$. Hence, the AUC can be seen as a sum of blocks of equal area of $\frac{1}{mn}$, showing that (7) holds.

Given the set of ROC curves for which the number of blocks under the curve is k, we can assign to each ROC curve a sequence $k_1, k_2, \ldots, k_r$ where k_1 is the number of blocks between $TMR = 0$ and $TMR = \frac{1}{m}$, until k_r, being the number of blocks between $TMR = \frac{r-1}{m}$ and $TMR = \frac{r}{m}$. By construction, (a) $k_1 + \ldots + k_r = k$, (b) the size of k_i is restricted to n, (c) r is limited to m, and (d) $k_1 \geq k_2 \geq \cdots k_r$.

The reverse relation also holds: given a sequence $k_1, k_2, \ldots, k_r$ with properties (a)-(d), we can construct the corresponding ROC curve uniquely as follows. Place k_1 blocks to the right between $TMR = 0$ and $TMR = \frac{1}{m}$, until k_r blocks to the right between $TMR = \frac{r-1}{m}$ and $TMR = \frac{r}{m}$.

The properties (a)-(d) of a sequence $k_1, k_2, \ldots, k_r$ make it a restricted partition of k. Since there is a one-to-one correspondence between a ROC curve and a restricted partition, we conclude that the number of ROC curves with $AUC = \frac{k}{mn}$ is equal to $p(n,m;k)$.

Given that the total number of ROC curves is $\binom{n+m}{n}$, all being equiprobable due to the same score distribution assumption, we conclude that (8) holds. □

We can also approximate (8) with the normal distribution.

Theorem 2. Given m genuine and n imposter scores, the distribution of the AUC has an asymptotic normal distribution if $m \to \infty$ and $n \to \infty$, in particular

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{p}(\mathrm{AUC} \geq x) = 1 - \Phi\left(\frac{(x-\frac{1}{2})mn}{\sigma_{mn}}\right). \tag{9}$$

Here Φ is the cumulative standard normal distribution and

$$\sigma_{mn} = \sqrt{\frac{mn(m+n+1)}{12}}. \tag{10}$$

Proof. According to Theorem 4 of [Ta86], we have, using our notation

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{p}\left(\frac{k-\frac{1}{2}mn}{\sigma_{mn}} \leq t\right) = \Phi(t), \tag{11}$$

with k related to AUC as $\mathrm{AUC} = \frac{k}{mn}$. Using this relation in (11) we observe that

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{p}\left(\frac{(\mathrm{AUC}-\frac{1}{2})mn}{\sigma_{mn}} \leq t\right) = \Phi(t), \tag{12}$$

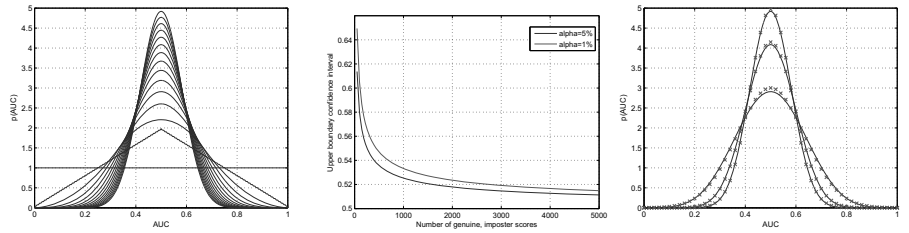
defining $x = \frac{1}{2} + \frac{t\sigma_{mn}}{mn}$ and reversing the inequality in (12) we conclude that (9) holds. □

5 Examples

In this section, we provide three examples of the exact distribution and one application that uses the approximation.

5.1 The 1 genuine/n imposter case

It is straightforward to show that $p(n,1;k) = \frac{(1-q)\cdots(1-q^{n+1})}{(1-q)\cdots(1-q^n)(1-q)} = \frac{1-q^{n+1}}{1-q} = \sum_{k=0}^{n} q^k$. Hence, $p(n,1;k) = 1$ for $k = 0,\ldots,n$. Moreover, $\mathrm{p}(\mathrm{AUC} = \frac{k}{mn}) = \frac{1}{\binom{n+1}{n}} = \frac{1}{n+1}$. This is in accordance with the example discussed in the Introduction.



a) b) c)

Fig. 1: a) p(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores. Graphs are scaled such that they can be interpreted as continuous probability distributions. b) The upper limit of 95% and 99% confidence intervals as a function of equal number of genuine and imposter scores. c) p(AUC) for $m = 5, 10, 15$ genuine and $n = 100$ imposter scores in blue, together with the approximation given by (9) in red.

5.2 The 2 genuine/Even n imposter case

Suppose n is even, then it can be shown that (5) can be written as

$$p(n,2;k) = (1+q+q^2+\cdots+q^n)(1+q^2+q^4+\cdots+q^n). \tag{13}$$

A straightforward calculation gives a staircase like shape:

$$\begin{aligned} &p(2,n;2k) = p(2,n;2k+1) = k+1 && \text{if } 2k \leq n-1, \\ &p(2,n,k) = \tfrac{n}{2}+1 && \text{if } k = n, \\ &p(2,n;k) = p(2,n;2n-k) && \text{if } k \geq n+1. \end{aligned} \tag{14}$$

5.3 The 1-15 genuine/100 imposter case

In this example, we plot p(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores in Figure 1a. In particular, we see respectively the uniform and staircase like shapes appearing for $m = 1$ and $m = 2$.

5.4 Confidence bounds

Theorem 2 can be used to construct a two sided $1-\alpha$ confidence interval $[\frac{1}{2}-x_\alpha, \frac{1}{2}+x_\alpha]$ around the AUC of a random classifier that depends on the number of genuine and imposter scores. Rewriting (9) shows that x_α is given by $x_\alpha = z_\alpha \sqrt{\frac{m+n+1}{12mn}}$, with z_α defined implicitly as $\Phi(z_\alpha) = 1 - \frac{\alpha}{2}$.

In Figure 1b we have chosen $m = n$, respectively $\alpha = 5\%$ ($z_\alpha = 1.96$) and $\alpha = 1\%$ ($z_\alpha = 2.33$) and plotted the upper limit of confidence intervals as a function of the number of genuine and imposter scores. This illustrates the asymptotic behaviour of the approximation; for smaller numbers of scores, the AUC of a random system can still deviate much from AUC=0.50.

6 Discussion

Figure 1a visualises the dependency of p(AUC) on the number of scores. Especially, we observe that for a lower number of scores, the probability that a random system has an AUC that differs significantly from 0.50 is non trivial. This is of relevance in, for example, the case of a subject anchored approach to evidence evaluation.

Although Theorem 1 provides an exact result, it can be challenging to calculate the value of the restricted partition function. One needs to resort to data structures to accommodate for values that are larger than those can fit into an IEEE-754 64 bit integer representation. This may result in an increased calculation time due to the lack of an efficient mapping from primitive operators to single machine instructions. Moreover, if we would be interested in the cumulative probability p(AUC $\geq x$), then a repeated calculation is not optimal as one could better use its generating function (4) for the simultaneous calculation of $p(n,m;k)$ over a range of values of k.

The result of Theorem 2 is an approximative result, and it is instructive to see how well it approximates the true probability distribution for finite values of m and n. We show the exact and the approximation for three cases: $m = 5,\ 10,\ 15$, and $n = 100$ in Figure 1c. Even for moderate values of m and n the approximation seems satisfactory. Furthermore, if the number of genuine and imposter scores are equal (k) and $k \to \infty$, the distribution becomes centered around AUC=0.50.

Although our work only considered approximative confidence bounds, we can also construct exact confidence bounds, especially when the number of scores is low.

7 Conclusion

In this paper, we have presented an exact formula for the probability distribution of the AUC of a random classifier, given a finite number of distinct genuine and imposter scores. This work can be used in the situation when we want to determine the probability that a random classifier produces the measured AUC; this is of interest when the measured AUC is low or the total number of scores is limited, masking the true nature of the classifier. The exact probability distribution involves the restricted partition function and can be approximated by a normal distribution. We used this approximation to derive confidence intervals for the AUC as a function of the number of genuine and imposter scores.

References

[An74] Andrews, George E.: Applications of basic hypergeometric functions. SIAM review, 16(4):441–484, 1974.

[An98] Andrews, G. E.: The Theory of Partitions. Cambridge Mathematical Library. Cambridge University Press, 1998.

[CM04] Cortes, Corinna; Mohri, Mehryar: Confidence intervals for the area under the ROC curve. In: Nips. pp. 305–312, 2004.

[DDCP88] DeLong, Elizabeth R.; DeLong, David M.; Clarke-Pearson, Daniel L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, pp. 837–845, 1988.

[Fa06] Fawcett, Tom: An introduction to ROC analysis. Pattern Recognition Letters, 27(8):861–874, 2006.

[Ha10] Hanczar, Blaise; Hua, Jianping; Sima, Chao; Weinstein, John; Bittner, Michael; Dougherty, Edward R.: Small-sample precision of ROC-related estimates. Bioinformatics, 26(6):822–830, 2010.

[HM82] Hanley, James A.; McNeil, Barbara J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1):29–36, 1982.

[HSZ09] Hsieh, Hsin-Neng; Su, Hsiu-Yuan; Zhou, Xiao-Hua: Interval estimation for the difference in paired areas under the ROC curves in the absence of a gold standard test. Statistics in medicine, 28(25):3108–3123, 2009.

[HT01] Hand, David J.; Till, Robert J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2):171–186, 2001.

[Me06] Meuwly, D.: Forensic Individualisation from Biometric Data. Science & Justice, 46(4):205–213, 2006.

[MG02] Mason, Simon J.; Graham, Nicholas E.: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society, 128(584):2145–2166, 2002.

[MHS98] Metz, Charles E.; Herman, Benjamin A.; Shen, Jong-Her: Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in medicine, 17(9):1033–1053, 1998.

[MW47] Mann, Henry B.; Whitney, Donald R.: On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, pp. 50–60, 1947.

[OL98] Obuchowski, Nancy A.; Lieber, Michael L.: Confidence intervals for the receiver operating characteristic area in studies with small samples. Academic Radiology, 5(8):561–571, 1998.

[PA95] Pham, T.; Almhana, J.: The generalized gamma distribution: its hazard rate and stress-strength model. IEEE Transactions on Reliability, 44(3):392–397, 1995.

[QH08] Qin, Gengsheng; Hotilovac, Lejla: Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Statistical Methods in Medical Research, 17(2):207–221, 2008.

[QZ06] Qin, Gengsheng; Zhou, Xiao-Hua: Empirical likelihood inference for the area under the ROC curve. Biometrics, 62(2):613–622, 2006.

[Ta86] Takács, Lajos: Some asymptotic formulas for lattice paths. Journal of Statistical Planning and Inference, 14(1):123–142, 1986.

[To77] Tong, Howell: On The Estimation of $\Pr\{Y < X\}$ for Exponential Families. IEEE Transactions on Reliability, 26(1):54–56, 1977.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 267

Fingerprint Damage Localizer and Detector of Skin Diseases from Fingerprint Images

Stepanka Barotova[1], Martin Drahansky[1]

Abstract: This article describes a novel approach for detection and classification of skin diseases in fingerprints using three methods - Block Orientation Field, Histogram Analysis and Flood Fill. The combination of these methods brings a surprising results and using a rule descriptor for selected skin diseases, we are able to classify the disease into a group or concrete name.

Keywords: Fingerprint recognition, skin diseases, image processing, image quality.

1 Introduction

Fingerprint-based systems are the most widely used biometric technology, which is very well accepted by users. Some people might use it literally on a daily basis, others only for civil identification or access systems. However, there is a significant number of people whose fingertip skin is affected of some kind of skin disease. Therefore, they cannot use fingerprint systems since skin diseases cause damages in ridge patterns.

The challenge now is to recognize the presence of skin diseases in fingerprint images and, if possible, eliminate their influence on the fingerprint recognition process, so that people suffering from skin diseases would be able to use fingerprint devices, at least to some extent.

Algorithms developed in our research are now able to locate the damage in the fingerprint image. Moreover, our classifier is able to estimate the possible disease present in the fingerprint image. This can have a great usage in forensic or medical applications, as well as security.

In this text, the methods used for localizing the damage and determining its type are going to be introduced, as well as the classification procedure and results.

2 The Triple-Method Damage Localization

There are three major methods that are used for the damage localization: *Block Orientation Field*, *Histogram Analysis* and *Flood Fill*. What makes the resulting concept so interesting, however, is their combination that provide valuable information about the quality and character of the possible disease.

[1] Faculty of Information Technology at Brno University of Technology, Department of Intelligent Systems, Bozetechova 2, 612 00 Brno, Czech Republic, xbarot00@stud.fit.vutbr.cz, drahan@fit.vutbr.cz

The classification is then based on the features extracted from the image by the Flood Fill algorithm, and their properties.

2.1 Ridge Inconsistence Detection from a Block Orientation Field

The computation of block orientation field is commonly used in the fingerprint recognition process for the purposes of estimating the ridges direction and classifying the fingerprint image into one of the several fingerprint classes [Ma09] [JFR08]. Because a typical fingerprint pattern consists of alternating dark and white lines, this information can be easily processed by a gradient operator that estimates the image gradient for each pixel. This low-level information is gathered and averaged for each $w \times w$ block in the image [HWJ98]. The transformation can result in a relatively smooth and continual image of the ridges direction estimates - for a healthy fingerprint of course - see Figure 1 on the left.

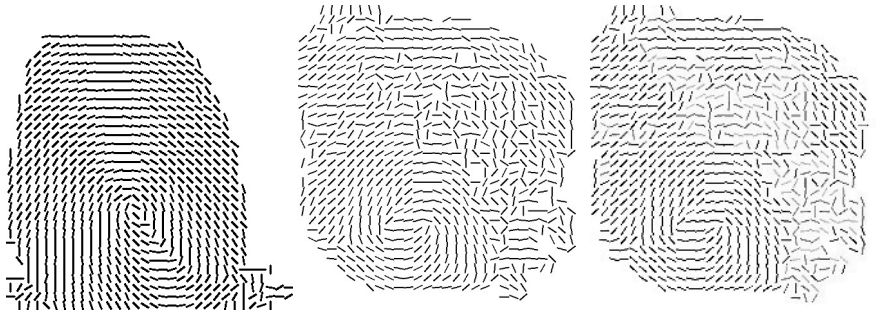


Fig. 1: Examples of block orientation images (left: healthy fingerprint, middle: fingerprint affected by a skin disease, right: detected damaged areas).

If we try to compute the block orientation field for a damaged or a partially damaged fingerprint however, the orientation field in damaged areas will be discontinuous, as displayed in Figure 1 in the middle. Exceptions to this are the peripheral areas and deltas and cores. These discontinuities can be detected by scanning the field for differences in direction angles.

The steps of the gradient-based method of block orientation field computation are as follows [HWJ98]:

1. Compute the gradients ∂_x and ∂_y for each pixel at (i, j) using a gradient operator. In this case a simple Sobel operator was used.
2. Divide the original image into $w \times w$ blocks.
3. Compute the estimation $\theta(i, j)$ of the ridge orientation for every image block centered at (i, j) using the Equations 1, 2 and 3:

$$v_x = \sum_{u=i-\frac{w}{2}}^{u=i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{v=j+\frac{w}{2}} 2\partial_x(u,v)\partial_y(u,v) \tag{1}$$

$$v_y = \sum_{u=i-\frac{w}{2}}^{u=i+\frac{w}{2}} \sum_{v=j-\frac{w}{2}}^{v=j+\frac{w}{2}} \partial_x^2(u,v)\partial_y^2(u,v) \tag{2}$$

$$\theta(i,j) = \frac{1}{2}\tan^{-1}\left(\frac{v_y(i,j)}{v_x(i,j)}\right) \tag{3}$$

The resulting block orientation field is afterwards analyzed for any discontinuities that may occur. The analysis is done using a row-wise and column-wise scanning approach that reveals areas of possible damage in the fingerprint. Neighboring blocks' directions are compared and a block is marked as a discontinuity if $|\theta(i,j) - \theta(i,j+1)| > 45^\circ$, where both estimations $\theta(i,j)$ and $\theta(i,j+1)$ have a value between 0° and 180°. Example detection is shown in Figure 1.

The advantage of this method is that it is already a part of the standard fingerprint recognition pipeline, so the algorithm can be easily implemented into existing methods. Also, it provides a fairly accurate estimate of the fingerprint damage in the sample.

2.2 Fingerprint Damage Detection using Histogram Analysis

This experimental method is based on the presumption that a quality fingerprint image consists of equally distributed ridges and valleys. If we assume that ridges are roughly the same dark color while valleys are light-colored, a histogram computed from each subfield of the fingerprint's area should ideally have a bimodal shape: it should have two peaks of approximately the same height and one valley between them, as displayed in Figure 2 on the left.

On the other hand, the intensity distribution in a fingerprint image part that belongs to a damaged area is not always as equal as in the quality one. Experiments showed that the majority of histograms computed from damaged subfields break the rules of the bimodal histogram. The lower the quality, the less the histogram resembles the ideal one. A non-bimodal histogram always implies a damaged or low-quality area, whereas a damaged subfield does not necessarily imply a non-bimodal histogram because a histogram is a measure for the distribution of intensities only and it does not take into account the pattern or neighborhoods of pixels. Figure 2 shows examples of non-bimodal histograms.

Fig. 2: Left: ideal bimodal histogram, others are examples of histograms computed from damaged areas.

The steps of the algorithm are as follows:

1. Divide the image into $w \times w$ blocks (ROIs = regions of interest), according to the desired resolution.
2. For each ROI, compute a histogram.
3. Check if the histogram is consistent with the bimodal characteristics generally found when finger ridges are healthy.
   a) Find all peaks and valleys of the histogram.
   b) If $peaks == 2$ and $valleys == 1$, histogram is bimodal, so continue with 3c. Otherwise quit: the histogram is non-bimodal.
   c) Check the heights and distances of the peaks and valleys. If the histogram passes these validity tests, it is bimodal, otherwise it is non-bimodal.

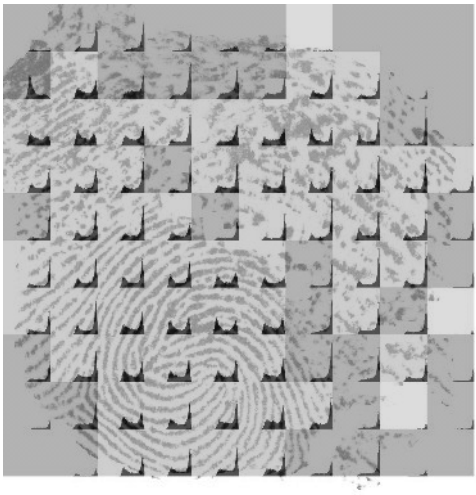


Fig. 3: Histogram Analysis result with details of particular histograms. Red background implies an invalid histogram, green means valid and blue stands for background.

Histogram Analysis is able to detect many areas the Block Orientation Field method might have omitted, therefore it is extremely valuable for the final determination of healthy areas.

Since the Histogram Analysis method is an experimental one, its results are not always accurate. Its drawback is the inability to cope with low-quality, especially dark, images. By implementing appropriate preprocessing steps, the method's performance and accuracy can be improved.

2.3 Features Extraction Based on the Flood Fill Algorithm

Flood Fill is a well known algorithm used for graphical purposes [GG08] and is especially handy for detecting and filling connected single-colored areas of an image. We have used

this characteristics to find local features of damaged fingerprints, such as straight lines or spots. These features are later used for the classification process.

In order to use the Flood Fill algorithm, the sample first needs to be preprocessed to obtain a black and white image. The preprocessing steps heavily depend on the image quality, as well as the type of sensor used for the acquisition. Therefore, for each database, they might differ. We have tailored the algorithm for our internal fingerprint database.

There are four types of features the Flood Fill algorithm is programmed to detect: large white spots, thick white lines, small dark spots and oblong dark lines (papillary lines disruptions). This is done using filtering the extracted areas according to specific parameters, such as the area's size or shape.

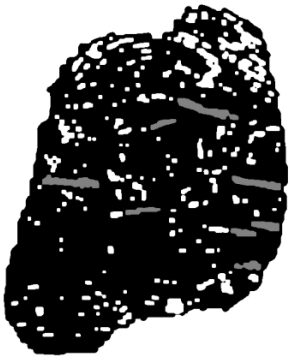


Fig. 4: Extraction of straight white lines.

2.4 Connection of the Methods

Connecting all three of the above-described methods together results in a surprisingly accurate description of the extent of damage in an entire area of a fingerprint image. They complement well as each of them detects a different kind of damage in the image.

At the end of each of the three detection methods, every image pixel is assigned a value 0 (healthy area), or a positive value up to 1 (damaged area). The greater the value, the more damaged the area to which the pixel belongs. Moreover, for the purpose of distinguishing fingerprint area from background, background was extracted separately according to [DH10] and the resulting information was stored in a fourth array with values -1 for (background) and 1 (fingerprint area).

The challenge was to connect these four output matrices together into a so-called Status Map which would give a good overview of the damage state every $w \times w$ block of pixels.

This is the description of the Status Map merging process:

1. Choose the resolution of the resulting Status Map.

2. Get the three output matrices and a background matrix.
3. For each matrix, compute a generalized block matrix (Status Map) that will store the average pixel values from $w \times w$ blocks: m_1, m_2, m_3 and *bckgr*.
4. Assign a weight to each method, according to the desired output and input image quality: w_1, w_2, w_3. Default values are: Orientation Field 2, Histogram Analysis 1 and Flood Fill 3.
5. For each block, compute its damage index. Damage index is a weighted mean of m_1, m_2 and m_3, masked by the value of the *bckgr* matrix.

$$damageIndex(i,j) = bckgr(i,j) * \frac{w_1 * m_1(i,j) + w_2 * m_2(i,j) + w_3 * m_3(i,j)}{w_1 + w_2 + w_3}$$

6. *damageIndex* now represents the extent of damage in each image block. The resulting Status Map gives an excellent overview of the damage.

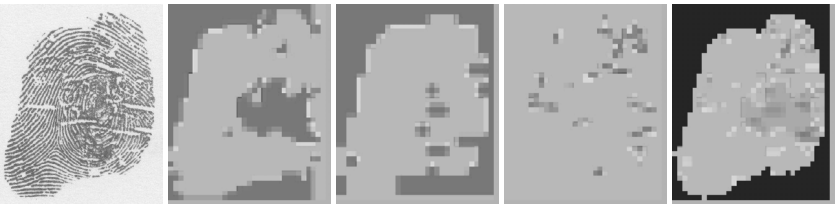


Fig. 5: Example of the pipeline of Status Maps and the final distribution of damage in the image (*atopic eczema*). Green color marks the healthy areas, blue color highlights the background and for the damaged areas a scale from yellow to red is used. Yellow stands for minor damage, whereas red implies extremely damaged places.

3 The Classification Process

The Classifier decides based on features extracted by the Flood Fill method and classifies the fingerprint image, according to the features' numbers and types. We have trained our classifier for 4 diseases: *acrodermatitis*, *atopic eczema*, *psoriasis* and *verruca vulgaris* [Ha09].

The decision rules have been determined with the help of statistics obtained from running the detector on our database of approximately 600 samples - see Table 1.

Tab. 1: Statistics of features extracted from each disease.

	acrodermatitis		atopic eczema		psoriasis		verruca vulgaris	
	med.	**std.dev.**	**med.**	**std.dev.**	**med.**	**std.dev.**	**med.**	**std.dev.**
white spots	5	3.97	5	4.31	8	5.35	1	3.02
white lines	2	1.84	3	3.06	4	2.65	1	1.63
dark spots	47	42.70	29	17.50	21	19.61	18	10.90
dark lines	7	8.37	17	19.80	8	9.22	15	39.76

Also, each disease has been given minimal value for some features (for example, *verruca vulgaris* logically has to have at least one white spot). All these characteristics are used in

order to compute an estimated likelihood that a certain set of features belong to a particular disease. The classifier chooses the disease with the highest likelihood.

4 Results

Thanks to the connection of the detection methods, very satisfactory results have been achieved for locating the damaged areas - as an example, see Figure 6. The classifier accuracy reached interesting values as well, as described below.

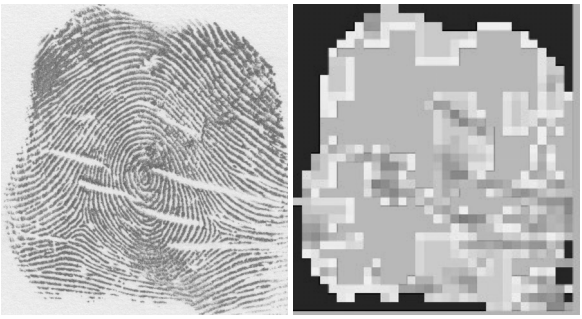


Fig. 6: Example of a final Status Map.

4.1 Classifier Accuracy

The classifier itself relies on the detection results. So far, the following accuracy measures have been computed for each disease class: FAR (*False Accept Rate*) and FRR (*False Reject Rate*) [Po11], ACC (total accuracy) - see Table 2. In this context, to *accept* means to classify a fingerprint into the disease class for which the measurements are being computed, whereas to *reject* means to classify a fingerprint into a different disease class. For the computation we used numbers of TP (*True Positives*), FN (*False Negatives*), FP (*False Positives*) and TN (*True Negatives*).

611 fingerprint images from dactyloscopic cards from the database were used for testing. The images had already been classified into disease classes by medical specialists. Table 2 shows the numbers of fingerprint images for each disease that were correctly/incorrectly classified by the algorithm.

Tab. 2: Classifier accuracy measures.

	TP	FN	FP	TN	FAR	FRR	ACC
Acrodermatitis	10	20	81	500	0.1394	0.6667	0.8347
Atopic eczema	126	297	37	151	0.1968	0.7021	0.4533
Psoriasis	31	87	168	325	0.3408	0.7373	0.5827
Verruca vulgaris	20	20	133	438	0.2329	0.5000	0.7496

The classification accuracy reached high values for for *acrodermatitis* (83.5%) and *verruca vulgaris* recognition (75.0%), whereas it was lower for *atopic eczema* (45.3%) and *psoriasis* (58.3%). The Classifier itself is ready to be further extended and improved.

5 Conclusion

We have developed algorithms that reach great quality in describing the overall extent of damage in a fingerprint image. The following methods were implemented: detection from block orientation field, Histogram Analysis method and an extended Flood Fill method. The best results were achieved by connecting the methods together using a Status Map. Along with the localizer, a classifier of four skin diseases was developed. It reached an accuracy of 83.5% for *acrodermatitis*, 45.3% for *atopic eczema*, 58.3% for *psoriasis* and 75.0% for *verruca vulgaris*.

There is a great potential for improvements and enhancements, and it is assumed that the research will continue. There are opportunities for the results of this research to be used in real-life applications in the future, such as medical applications or programs for police and security purposes.

Acknowledgement

This work was supported by The Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability (NPU II); project IT4Innovations excellence in science - LQ1602. and the university project Secure and Reliable Computer Systems FIT-S-17-4014.

References

[DH10] Dolezel, Michal; Hejtmankova, Dana: Segmentation Procedure for Fingerprint Area Detection in Image Based on Enhanced Gabor Filtering. International Journal of Bio-Science and Bio-Technology, 2(4):39–50, December 2010.

[GG08] Godse, D. A.; Godse, A. P.: Computer Graphics. Technical Publications, 2008.

[Ha09] Habif, Thomas P.: Clinical Dermatology. Edinburgh: Mosby, 5 edition, 2009.

[HWJ98] Hong, Lin; Wan, Yifei; Jain, A.: Fingerprint image enhancement: algorithm and performance evaluation. IEEE Transactions Pattern Analysis and Machine Intelligence, 20(8):777–789, 1998.

[JFR08] Jain, A. K.; Flynn, P.; Ross, A. A.: Handbook of Biometrics. Springer-Verlag, 2008.

[Ma09] Maltoni, Davide; Maio, Dario; Jain, Anil K.; Prabhakar, Salil: Handbook of Fingerprint Recognition. Springer, 2 edition, 2009.

[Po11] Powers, D. M. W.: Evaluation: From Precision, Recall and F-measure to ROC, Informedness & Correlation. Journal of Machine Learning Technologies, 2:37–63, 2011.

A. Brömme, C. Busch, A. Dantcheva, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2017,
Lecture Notes in Informatics (LNI), Gesellschaft fur Informatik, Bonn 2017 275

Fusing Biometric Scores using Subjective Logic for Gait Recognition on Smartphone

Pankaj Wasnik*‡, Kirstina Schäfer†‡, Kiran Raja*, Raghavendra Ramachandra*, C. Busch*

Abstract: The performance of a biometric system gets affected by various types of errors such as systematic errors, random errors, etc. These kinds of errors usually occur due to the natural variations in the biometric traits of subjects, different testing, and comparison methodologies. Neither of these errors can be easily quantifiable by mathematical formulas. This behavior introduces an uncertainty in the biometric verification or identification scores. The combination of comparison scores from different comparators or combination of multiple biometric modalities could be a better approach for improving the overall recognition performance of a biometric system. In this paper, we propose a method for combining such scores from multiple comparators using *Subjective Logic (SL)*, as it takes uncertainty into account while performing to biometric fusion. This paper proposes a framework for a smartphone based gait recognition system with application of *SL* for biometric data fusion.

Keywords: subjective logic, biometric score fusion, gait recognition, smartphone biometrics, user verification, pattern recognition

1 Introduction

Gait, the walking manner of a person, can be used to distinguish between individuals. By placing an accelerometer sensor on the body of a person, the recorded signal can be used to identify that person. Commercial mobile phones nowadays have accelerometer sensors included as a standard feature and can be easily used for gait recognition. Hence, this makes gait recognition a viable alternative to other traditional means such as password or lock patterns for validating a user for phone's ownership or any other high security demanding applications such as online banking, etc. The password or lock patterns, typically have to be remembered by the user and given manually upon prompting of the phone. The gait, on the other hand, can be observed unobtrusively while the phone is inside the trouser pocket. It cannot be lost or forgotten, due to it being a behavioral characteristic of an individual.

The technical report ISO/IEC 19795-1 describes the performance of a biometric system and errors related to them. It explains how a biometric system performance is affected by systematic and random errors. These types of errors can occur due to the natural variations in the biometric traits of subjects, different testing and comparison methodologies and this brings an uncertainty in the biometric recognition[IS06]. One of the reasonable approaches to deal with such errors is Biometric fusion. The biometric fusion can happen at different levels, and one of such methods is score-level fusion which is combining the comparison scores of various comparators to improve the biometric performance[Ul06]. In recent years, researchers have proposed several fusion strategies. But none of them take

* Norwegian Biometrics Laboratory, NTNU, Gjøvik, Norway
{pankaj.wasnik;raghavendra.ramachandra;christoph.busch;kiran.raja}@ntnu.no
† Faculty of Computer Science, Hochschule Darmstadt, Germany, kristina_schaefer@gmx.net
‡ These authors have contributed equally to this article

the uncertainty of biometric systems into account, which leads to ignorance towards the performance characteristics of the biometric system under consideration. The systematic and random errors aren't quantifiable easily, yet we need to estimate an uncertainty of the system. The technical report ISO/IEC 19795-1 [IS06] provides some methods to estimate it.

In this paper, we propose a fusion method based on *SL* which takes into account an uncertainty of the system. Hence, limitations of the biometric performance can also be considered while performing the score level fusion. Further, the remaining paper is mainly divided into Related Work, Proposed Method, Experiments & Evaluation followed by Conclusions.

2 Related Work

One of the earliest studies into gait recognition using wearable accelerometers was published by Mantyjarvi in 2005 [Ma05]. The author proposed a technique based on distances between two extracted steps using matching pattern techniques. Further advances were made by Gafurov [GSB07] and Derawi [DBH10] proposing optimizations for cycle detection. Derawi and Bours and Shrestha [BS] also improved methods to calculate the distance between two cycles. Further, Nickel [Ni11] [NWB12] and Watanabe [WS] proposed methods using fix-length segments instead of extracted cycles. [ZD14] [ZDM15] focus their work on creating a gait recognition system which is not dependent on the subjects walking pace and the orientation of the accelerometer sensor. [MPM16] proposes a normalization procedure for cross-device gait recognition.

Subjective Logic was first introduced by Jøsang[Jø97] as an extension of probability calculus and binary logic. It operates on subjective beliefs to serve an opinion about whether the world is true or false. The term opinion represents the subjective belief. Subjective logic operates on these opinions and also contains standard and non-standard logical operators[Jø]. Further, Jøsang[Jø] described opinions could be interpreted as a probability measure providing secondary uncertainty. The application of subjective logic in the domain of biometrics was recently introduced by Jøsang et al. in [JMM14] where authors have described the use of various Subjective Logic Fusion (SLF) operators in biometric fusion via belief fusion to produce a new opinion by fusing opinions from different sources.

In this paper, we are fusing comparison score opinions from different comparators using three *SLF* operators i.e. averaging, cumulative and consensus. These types of fusion operators are the common choice for situations like consistent and inconsistent score opinions. Here, in this case, it is very much possible as three classifiers will have different scores based on their recognition performance and also consist of uncertainty in the output scores.

3 Proposed Fusion Scheme

Here, we first describe the various terms involved in the proposed scheme related to the subjective logic. As per best of our knowledge, this is the first work which is trying to simplify and check the feasibility of the models proposed in [JMM14] for real world application and evaluation of these methods to verify the applicability and usability of *SLF* in biometrics. The subsequent sections describe each term in details as follows:

3.1 Subjective Logic

The primary objective of *SL* is to enhance probabilistic logic by including uncertainty about input probabilities and introducing subjective belief in these probabilities [Jø11]. It combines the probabilistic logic, uncertainty, and subjectivity to form a firm opinion. *"Arguments in subjective logic are called subjective opinions, or opinions for short. An opinion can contain uncertainty mass in the sense of uncertainty about probabilities[Jø11]"*. The biometric similarity scores i.e. whether the user is a genuine or impostor can simply be expressed as *a binomial opinion*. Consider binary domain $X = \{x, \bar{x}\}$ where x is a binary variable representing an user being genuine and $\bar{x}$ is the compliment of x i.e. subject is not a genuine user. Furthermore, binomial subjective opinion about a person being genuine user can be represented by quadruple $\omega_X = (b_x^{c_i}, d_x^{c_i}, u_x^{c_i}, a_x^{c_i})$ where, b_x, d_x, u_x *and* a_x are classifier c_i's belief, disbelief, uncertainty about probability of x and base rate or prior probability of X respectively. Here, $i = 1, 2...n$, where n is the number of classifiers. For any given subjective opinion $\omega = (b, d, u, a)$, ***Belief Subjective Additivity*** theorem is always true[Jø11] and is expressed by Eq. 1:

$$b_x^c + d_x^c + u_x^c = 1 \tag{1}$$

For binomial opinions, the projected probability of x can be expressed as defined by Eq. 2

$$\omega(x) = b_x + u_x a_x \tag{2}$$

We adopt this knowledge to formulate the proposed scheme for biometric score fusion by transforming the biometric similarity scores into the subjective opinions. These subjective opinions are later used in subjective logic fusion. From here onwards, we assume the comparison scores as belief(b), the prior probability of the subject being genuine or impostor as base rate(a) and $\hat{V}(\hat{p})$ as an uncertainty(u) which is described in the biometric standards ISO/IEC 19795 Part-1 [IS06] and given by Eq. 3

$$\hat{V}(\hat{p}) = \frac{\sum_{i=1}^{n} a_i^2 - 2\hat{p}\sum_{i=1}^{n} a_i m_i + \hat{p}^2 + \sum_{i=1}^{n} m_i^2}{\frac{n-1}{n}\sum_{i=1}^{n} m_i}, \; where \; \hat{p} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} m_i} \tag{3}$$

where, n is the number of enrolled test subjects, m_i is the number of attempts by i^{th} subject, a_i is the number of false-non matches for i^{th} subject. Therefore, for every comparison score S_i we will have a subjective opinion ω_i which is defined by a quadruple (b_i, d_i, u_i, a_i).

3.2 Proposed Scheme

This section describes an overview of the proposed fusion scheme using subjective logic. Figure 1 illustrates the overview of the proposed scheme in details. This study considers three well-known classifiers as comparators for obtaining the similarity scores. For analysis, we have used *Extremely Randomized Trees (ERT)*, *Multi-layer Perceptron (MLP)* and *Random Forest Classifier (RFC)* as our baseline comparators. ERT and RFC have a maximum 100 random trees in the forest while MLP has two hidden layers with 10 and 5 nodes, along with the length of the feature vector as the number of input nodes and 2 output nodes.

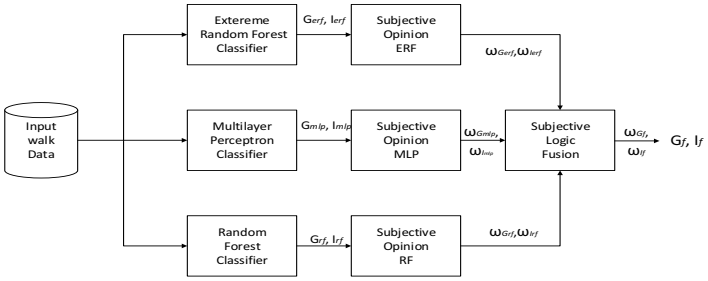


Figure 1: Proposed fusion scheme

All of the classifiers produce an output between $[0,1]$ where the maximum score represents 100% genuine subject and minimum score as 100% impostor. From the Figure 1 we can briefly understand the steps involved in the fusion process. Firstly, we pass the input data to baseline comparators i.e to ERT, MLP and RFC, their output is then processed to generate corresponding subjective opinions. These subjective opinions are fused using three *SLF* operators which are *1. Subjective Average 2. Consensus and 3. Cumulative Fusion*. In the last step, fused genuine and impostor opinions are used to obtain the corresponding scores through Equation 2.

Following subsections describe the fusion operators in detail

1. Averaging Fusion: Averaging opinion fusion is used when dependence between arguments from different observers are assumed as they will represent better observation together [JMM14]. Let ω^A *and* ω^B be the subjective opinions from source A and source B, then the fused opinion is $\omega^{A\diamond B} = \omega^A \underline{\oplus} \omega^B$, such that :

$$\begin{cases} b^{A\diamond B} = \frac{b^A u^B + b^B u^A}{u^A + u^B} & \forall\, u^A \neq 0 \text{ and } u^B \neq 0 \\ u^{A\diamond B} = \frac{2u^A u^B}{u^A + u^B} \end{cases} \tag{4}$$

$$\begin{cases} b^{A\diamond B}(x_i) = \gamma^A b^A(x_i) + \gamma^B b^B(x_i) & \forall\, u^A = 0 \text{ and } u^B = 0 \\ u_{A\diamond B} = 0 \end{cases} \tag{5}$$

where, $\gamma^A = \lim_{\substack{u^A\to 0 \\ u^B\to 0}} \frac{u^B}{u^A+u^B}$, $\gamma^B = \lim_{\substack{u^A\to 0 \\ u^B\to 0}} \frac{u^B}{u^A+u^B}$

2. Consensus Fusion: Consensus opinion fusion assumes that the input opinions are independent and combining them would reduce the uncertainty among them. Let ω^A *and* ω^B be the subjective opinions from source A and source B, then the fused opinion is $\omega^{A,B} = \omega^A \otimes \omega^B$, such that :

$$\begin{cases} b^{A\diamond B} = \frac{b^A u^B + b^B u^A}{u^A + u^B - u^A u^B} \\ u^{A\diamond B} = \frac{u^A u^B}{u^A + u^B - u^A u^B} \\ a^{A\diamond B} = \frac{a^A u^B + a^B u^A - (a^A a^B) u^A u^B}{u^A + u^B - u^A u^B} \end{cases} \tag{6}$$

3. Cumulative Fusion: Cumulative opinion fusion is used when we can increase the amount of evidence by including more arguments and the certainty increases with an increase amounting to evidence. Let ω^A *and* ω^B be the subjective opinions from source A and source B, then the fused opinion is $\omega^{A,B} = \omega^A \oplus \omega^B$, such that :

$$\begin{cases} b^{A\diamond B} = \frac{b^A u^B + b^B u^A}{u^A + u^B - u^A u^B} \\ u^{A\diamond B} = \frac{u^A u^B}{u^A + u^B - u^A u^B} \end{cases} \tag{7}$$

$$\begin{cases} b^{A\diamond B}(x_i) = \gamma^A b^A(x_i) + \gamma^B b^B(x_i) \quad \forall\, u^A = 0 \; and \; u^B = 0 \\ u_{A\diamond B} = 0 \end{cases} \tag{8}$$

where, $\gamma^A = \lim_{\substack{u^A \to 0 \\ u^B \to 0}} \frac{u^B}{u^A + u^B}$, $\gamma^B = \lim_{\substack{u^A \to 0 \\ u^B \to 0}} \frac{u^B}{u^A + u^B}$

In the above equations, the variables $b^{A\diamond B}, u^{A\diamond B}$ *and* $a^{A\diamond B}$ represent the fused belief, uncertainty and base rates satisfying Equation 1.

4 Database, Experiments and Evaluation

This section describes statistics of the database evaluated, the experiments carried out and obtained results in detail.

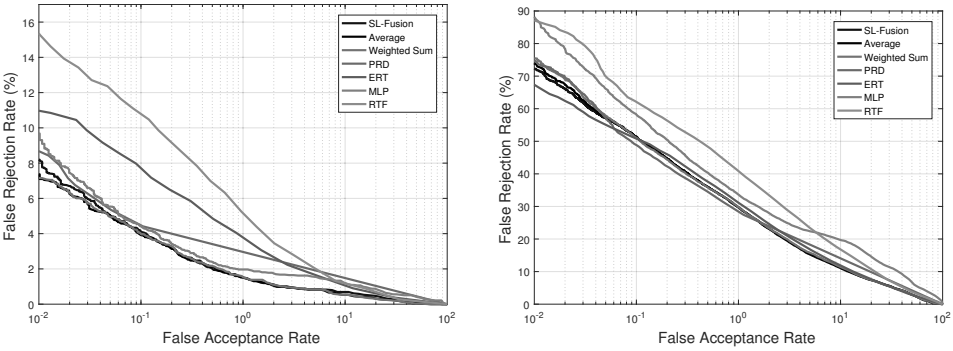
4.1 Database

The database which we used for evaluation is the previously collected database by Nickel [Ni12]. This database consists of 48 subjects, each with 2 walking sessions. The data was recorded using a smart-phone which was put inside a pouch fastened on the right side of the hip of the subject. The route was divided into 9 points to simulate realistic scenarios and data between start-point and end-point is considered as one whole walk. Three such walks were recorded per subjects. Thus, the database contains 27 samples from 9 different points along with two enrollment samples, and it consists of 2784 samples in total. In the evaluation three training strategies were formed i.e. *set1*, *set2* and *set3*. In each of these settings 9 samples (for example walk 1) are used as training samples, while samples of the remaining two walks (18 samples) are used for testing strategy 1. This setting is used for the remaining two strategies with corresponding changes. Hence, for each training set, we have 432 training samples from 48 subjects and 2160 testing samples (864 and 1296 of Session 1 and 2 respectively).

4.2 Experiments

We first do preprocessing and feature extraction. The steps involved are based on the work [Ni12]. As a first step the walk files were cleaned, if necessary, so that the data only included walking periods. As a second step, interpolation, potential irregular time intervals between the signals were corrected for a pre-defined frequency. Lastly, we normalize the data around zero to compensate for calibration irregularities of the accelerometer sensor. Next, we segment the walk data with Fix-length segmentation, which is achieved by dividing the data into equal parts of fixed length with an overlapping factor of 50%. The features

considered are statistical (ST), the histogram of the distribution (BIN), Mel-frequency cepstral coefficients (MF) and Bark-frequency cepstral coefficients (BF1 and BF2). For the classification, we used the best-performing features of all accelerometer axes data. When we concatenate all of these features together, the best results were achieved (See Table 1). **.



(a) DET for set1 data

(b) DET for set2 data

Figure 2: DET Curves for Session 1 data**

4.3 Evaluation

For the evaluation, we first identify the best performing feature set by executing various tests as discussed earlier. Table 1 represents the details of EERs for each feature extraction technique for all classifiers and training sets. The combination of features gives a better performance than if we use them separately (Ref. Table 1). Further, MLP classifier with *set1* training data gives the lowest EER of 1.77%. Next, we generate the comparison scores using ERT, MLP, and RFC. These scores are further processed to obtain the fused scores for testing data from Session 1 and 2. Table 2 presents the details of EER for both sessions. The presented results compare the proposed fusion scheme against the baseline and basic fusion strategies such as *average, weighted sum, and product rule*. From Table 2 we can observe, EER range differs a lot between Session 1 and 2. One of the possible reasons could be a change in the characteristics of the testing data due to the time gap between two session item. The proposed method outperforms all of the basic fusion techniques. We achieved an EER of **1.31%** and **9.96%** for Session 1 and 2 respectively using the proposed scheme. In both of the sessions, the *SLF* cumulative fusion has consistent performance i.e. it has the lowest EER for all tests except for *set3* and Session 1 testing data. The performance of *SL* averaging and cumulative fusion is nearly equal which signifies that increasing the evidence increased the performance of the system.

Furthermore, we analyze the Detection Error Trade-off (DET) curves to understand the operating characteristics of the proposed scheme. DET often plots False Rejection Rate (FRR) on the y-axis and False Acceptance Rate (FAR) on the x-axis in a logarithmic scale. As y-axis represents the number of match error, the curve close to the origin corresponds

** For simplicity, in the Figure 2, the SL-Fusion curve represents only the best performing fusion operation, which is found as the SL-Cumulative Fusion operator

Features	Set1			Set2			Set3		
	ERT	**MLP**	**RFC**	**ERT**	**MLP**	**RFC**	**ERT**	**MLP**	**RFC**
BF1BF2	2.75	2.73	3.62	3.40	3.38	4.14	3.24	2.31	4.09
BF1BIN5ST	2.87	3.49	3.49	3.36	3.60	4.20	3.15	3.38	3.88
BF2BIN5ST	3.35	3.53	3.64	3.68	3.95	4.20	3.42	3.20	4.02
MFBF1	3.09	2.90	3.64	3.43	2.92	4.24	3.29	2.72	4.01
MFBF2	3.19	3.36	3.89	3.60	3.52	4.21	3.56	3.46	4.34
MFBIN5ST	3.20	3.74	3.82	3.99	4.54	4.30	3.42	3.89	4.35
ALL	**2.40**	**1.77**	**2.94**	**2.98**	**2.32**	**3.57**	**2.99**	**2.93**	**3.51**

Table 1: EER for different feature sets

Comparators	Session 1			Session 2		
	set1	***set2***	***set3***	***set1***	***set2***	***set3***
ERT	2.40	2.98	2.99	11.64	10.01	11.56
MPL	1.77	2.32	2.93	16.44	13.11	15.64
RFC	2.94	3.57	3.51	13.71	13.09	12.68
SL average	1.34	1.76	**2.00**	10.97	9.99	11.02
SL cumulative	**1.31**	**1.71**	2.25	**10.66**	**9.96**	**10.91**
SL consensus	1.53	2.17	2.75	13.99	12.23	13.05
Average	1.34	1.79	2.18	10.73	10.01	10.95
Weighted Sum	1.38	1.85	2.05	10.88	10.01	10.92
Product	2.25	2.58	2.39	14.27	12.88	12.24

Table 2: EERs for Nikel's database[NWB12] and all features combined together

to the best performance. Due to the page limitation, we have presented the DET curves for Session 1 data with strategies *set1 and set2* (See Figure 2). Therefore, from Figure 2 we can see, the performance of the proposed scheme is higher than the individual classifiers. For the proposed scheme we achieved an FRR of 8.18% at FAR 1/100 with an EER of 1.31% for the testing data from Session 1. Further, we have obtained an average FRR of 64% approximately at FAR 1/100 and the lowest average EER of 10.5% across all training sets using the proposed scheme for Session 2 data.

5 Conclusion

We observed that the performance of baseline classifiers is worse than all of the *SLF* operators. For this challenging database, we have achieved an EER of 1.31% which is much less than the reported EER of 6.13% by Nickel[Ni12]. The proposed fusion scheme using subjective logic considers the errors of the biometric system when performing the fusion while other mentioned general biometric fusion methods ignore them. We have achieved the best results for *SL* cumulative fusion operator in terms of EER. We obtained lower EERs compared to the performance of the individual classifiers for *SLF* operators such as average and consensus. In conclusion, this paper successfully models biometric fusion to the *Subjective Logic Fusion* and proposes a simplified methodology for using it. For the fu-

ture work, different experiments & techniques could be explored to model the uncertainty and errors in the system to improvise fusion strategies to achieve higher performance.

6 Acknowledgment

This work was carried out under the funding from the Research Council of Norway (Grant No. IKTPLUSS 248030/O70).

References

[BS] Bours, P.; Shrestha, R.: Eigensteps: A giant leap for gait recognition. pp. 1–3.

[DBH10] Derawi, M. O.; Bours, P.; Holien, K.: Improved Cycle Detection for Accelerometer Based Gait Authentication. pp. 312–317, 2010.

[GSB07] Gafurov, D.; Snekkenes, E.; Bours, P.: Gait Authentication and Identification Using Wearable Accelerometer Sensor. IEEE, pp. 220–225, 2007.

[IS06] ISO/IEC: ISO/IEC IS 19795-1 Information Technology - Biometric performance testing and reporting- Part 1: Principles and framework. ISO/IEC, 2006.

[JMM14] Jøsang, Audun; Munch-Møller, Thorvald H: Biometric data fusion based on subjective logic. IEEE, pp. 1–8, 2014.

[Jø] Jøsang, Audun: A logic for uncertain probabilities.

[Jø97] Jøsang, Audun: Artificial reasoning with subjective logic. Perth:[sn], 1997.

[Jø11] Jøsang, Audun: Subjective logic. Book draft, 2011.

[Ma05] Mantyjarvi, J.; Lindholm, M.; Vildjiounaite, E.; Makela, S.; Ailisto, H.: Identifying Users of Portable Devices from Gait Pattern with Accelerometers. pp. 973–976, 2005.

[MPM16] Marsico, M. De; Pasquale, D. De; Mecca, A.: Embedded accelerometer signal normalization for cross-device gait recognition. IEEE, pp. 1–5, 2016.

[Ni11] Nickel, C.; Busch, C.; Rangarajan, S.; Mobius, M.: Using Hidden Markov Models for accelerometer-based biometric gait recognition. pp. 58–63, 2011.

[Ni12] Nickel, C.: Accelerometer-based Biometric Gait Recognition for Authentication on Smartphones. Ph.D. Thesis, Darmstadt, 2012.

[NWB12] Nickel, C.; Wirtl, T.; Busch, C.: Authentication of Smartphone Users Based on the Way They Walk Using k-NN Algorithm. pp. 16–20, 2012.

[Ul06] Ulery, Brad; Hicklin, Austin; Watson, Craig; Fellner, William; Hallinan, Peter: Studies of biometric fusion. 7346, 2006.

[WS] Watanabe, Y.; Sara, S.: Toward an Immunity-based Gait Recognition on Smart Phone: A Study of Feature Selection and Walking State Classification.

[ZD14] Zhong, Y.; Deng, Y.: Sensor orientation invariant mobile gait biometrics. pp. 1–8, 2014.

[ZDM15] Zhong, Y.; Deng, Y.; Meltzner, G.: Pace independent mobile gait biometrics. pp. 1–8, 2015.

GI-Edition Lecture Notes in Informatics

P-1 Gregor Engels, Andreas Oberweis, Albert Zündorf (Hrsg.): Modellierung 2001.

P-2 Mikhail Godlevsky, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications, ISTA'2001.

P-3 Ana M. Moreno, Reind P. van de Riet (Hrsg.): Applications of Natural Lan-guage to Information Systems, NLDB'2001.

P-4 H. Wörn, J. Mühling, C. Vahl, H.-P. Meinzer (Hrsg.): Rechner- und sensorgestützte Chirurgie; Workshop des SFB 414.

P-5 Andy Schürr (Hg.): OMER – Object-Oriented Modeling of Embedded Real-Time Systems.

P-6 Hans-Jürgen Appelrath, Rolf Beyer, Uwe Marquardt, Heinrich C. Mayr, Claudia Steinberger (Hrsg.): Unternehmen Hochschule, UH'2001.

P-7 Andy Evans, Robert France, Ana Moreira, Bernhard Rumpe (Hrsg.): Practical UML-Based Rigorous Development Methods – Countering or Integrating the extremists, pUML'2001.

P-8 Reinhard Keil-Slawik, Johannes Magenheim (Hrsg.): Informatikunterricht und Medienbildung, INFOS'2001.

P-9 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Innovative Anwendungen in Kommunikationsnetzen, 15. DFN Arbeitstagung.

P-10 Mirjam Minor, Steffen Staab (Hrsg.): 1st German Workshop on Experience Management: Sharing Experiences about the Sharing Experience.

P-11 Michael Weber, Frank Kargl (Hrsg.): Mobile Ad-Hoc Netzwerke, WMAN 2002.

P-12 Martin Glinz, Günther Müller-Luschnat (Hrsg.): Modellierung 2002.

P-13 Jan von Knop, Peter Schirmbacher and Viljan Mahni_ (Hrsg.): The Changing Universities – The Role of Technology.

P-14 Robert Tolksdorf, Rainer Eckstein (Hrsg.): XML-Technologien für das Semantic Web – XSW 2002.

P-15 Hans-Bernd Bludau, Andreas Koop (Hrsg.): Mobile Computing in Medicine.

P-16 J. Felix Hampe, Gerhard Schwabe (Hrsg.): Mobile and Collaborative Business 2002.

P-17 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Zukunft der Netze –Die Verletzbarkeit meistern, 16. DFN Arbeitstagung.

P-18 Elmar J. Sinz, Markus Plaha (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2002.

P-19 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund.

P-20 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund (Ergänzungsband).

P-21 Jörg Desel, Mathias Weske (Hrsg.): Promise 2002: Prozessorientierte Methoden und Werkzeuge für die Entwicklung von Informationssystemen.

P-22 Sigrid Schubert, Johannes Magenheim, Peter Hubwieser, Torsten Brinda (Hrsg.): Forschungsbeiträge zur "Didaktik der Informatik" – Theorie, Praxis, Evaluation.

P-23 Thorsten Spitta, Jens Borchers, Harry M. Sneed (Hrsg.): Software Management 2002 – Fortschritt durch Beständigkeit

P-24 Rainer Eckstein, Robert Tolksdorf (Hrsg.): XMIDX 2003 – XML-Technologien für Middleware – Middleware für XML-Anwendungen

P-25 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Commerce – Anwendungen und Perspektiven – 3. Workshop Mobile Commerce, Universität Augsburg, 04.02.2003

P-26 Gerhard Weikum, Harald Schöning, Erhard Rahm (Hrsg.): BTW 2003: Datenbanksysteme für Business, Technologie und Web

P-27 Michael Kroll, Hans-Gerd Lipinski, Kay Melzer (Hrsg.): Mobiles Computing in der Medizin

P-28 Ulrich Reimer, Andreas Abecker, Steffen Staab, Gerd Stumme (Hrsg.): WM 2003: Professionelles Wissensmanagement – Er-fahrungen und Visionen

P-29 Antje Düsterhöft, Bernhard Thalheim (Eds.): NLDB'2003: Natural Language Processing and Information Systems

P-30 Mikhail Godlevsky, Stephen Liddle, Heinrich C. Mayr (Eds.): Information Systems Technology and its Applications

P-31 Arslan Brömme, Christoph Busch (Eds.): BIOSIG 2003: Biometrics and Electronic Signatures

P-32 Peter Hubwieser (Hrsg.): Informatische Fachkonzepte im Unterricht – INFOS 2003

P-33 Andreas Geyer-Schulz, Alfred Taudes (Hrsg.): Informationswirtschaft: Ein Sektor mit Zukunft

P-34 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 1)

P-35 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 2)

P-36 Rüdiger Grimm, Hubert B. Keller, Kai Rannenberg (Hrsg.): Informatik 2003 – Mit Sicherheit Informatik

P-37 Arndt Bode, Jörg Desel, Sabine Rathmayer, Martin Wessner (Hrsg.): DeLFI 2003: e-Learning Fachtagung Informatik

P-38 E.J. Sinz, M. Plaha, P. Neckel (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2003

P-39 Jens Nedon, Sandra Frings, Oliver Göbel (Hrsg.): IT-Incident Management & IT-Forensics – IMF 2003

P-40 Michael Rebstock (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2004

P-41 Uwe Brinkschulte, Jürgen Becker, Dietmar Fey, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle, Thomas Runkler (Edts.): ARCS 2004 – Organic and Pervasive Computing

P-42 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Economy – Transaktionen und Prozesse, Anwendungen und Dienste

P-43 Birgitta König-Ries, Michael Klein, Philipp Obreiter (Hrsg.): Persistance, Scalability, Transactions – Database Mechanisms for Mobile Applications

P-44 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): Security, E-Learning. E-Services

P-45 Bernhard Rumpe, Wofgang Hesse (Hrsg.): Modellierung 2004

P-46 Ulrich Flegel, Michael Meier (Hrsg.): Detection of Intrusions of Malware & Vulnerability Assessment

P-47 Alexander Prosser, Robert Krimmer (Hrsg.): Electronic Voting in Europe – Technology, Law, Politics and Society

P-48 Anatoly Doroshenko, Terry Halpin, Stephen W. Liddle, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications

P-49 G. Schiefer, P. Wagner, M. Morgenstern, U. Rickert (Hrsg.): Integration und Datensicherheit – Anforderungen, Konflikte und Perspektiven

P-50 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 1) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-51 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 2) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-52 Gregor Engels, Silke Seehusen (Hrsg.): DELFI 2004 – Tagungsband der 2. e-Learning Fachtagung Informatik

P-53 Robert Giegerich, Jens Stoye (Hrsg.): German Conference on Bioinformatics – GCB 2004

P-54 Jens Borchers, Ralf Kneuper (Hrsg.): Softwaremanagement 2004 – Outsourcing und Integration

P-55 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): E-Science und Grid Adhoc-Netze Medienintegration

P-56 Fernand Feltz, Andreas Oberweis, Benoit Otjacques (Hrsg.): EMISA 2004 – Informationssysteme im E-Business und E-Government

P-57 Klaus Turowski (Hrsg.): Architekturen, Komponenten, Anwendungen

P-58 Sami Beydeda, Volker Gruhn, Johannes Mayer, Ralf Reussner, Franz Schweiggert (Hrsg.): Testing of Component-Based Systems and Software Quality

P-59 J. Felix Hampe, Franz Lehner, Key Pousttchi, Kai Ranneberg, Klaus Turowski (Hrsg.): Mobile Business – Processes, Platforms, Payments

P-60 Steffen Friedrich (Hrsg.): Unterrichtskonzepte für inforrmatische Bildung

P-61 Paul Müller, Reinhard Gotzhein, Jens B. Schmitt (Hrsg.): Kommunikation in verteilten Systemen

P-62 Federrath, Hannes (Hrsg.): „Sicherheit 2005“ – Sicherheit – Schutz und Zuverlässigkeit

P-63 Roland Kaschek, Heinrich C. Mayr, Stephen Liddle (Hrsg.): Information Systems – Technology and ist Applications

P-64 Peter Liggesmeyer, Klaus Pohl, Michael Goedicke (Hrsg.): Software Engineering 2005

P-65 Gottfried Vossen, Frank Leymann, Peter Lockemann, Wolffried Stucky (Hrsg.): Datenbanksysteme in Business, Technologie und Web

P-66 Jörg M. Haake, Ulrike Lucke, Djamshid Tavangarian (Hrsg.): DeLFI 2005: 3. deutsche e-Learning Fachtagung Informatik

P-67 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 1)

P-68 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 2)

P-69 Robert Hirschfeld, Ryszard Kowalcyk, Andreas Polze, Matthias Weske (Hrsg.): NODe 2005, GSEM 2005

P-70 Klaus Turowski, Johannes-Maria Zaha (Hrsg.): Component-oriented Enterprise Application (COAE 2005)

P-71 Andrew Torda, Stefan Kurz, Matthias Rarey (Hrsg.): German Conference on Bioinformatics 2005

P-72 Klaus P. Jantke, Klaus-Peter Fähnrich, Wolfgang S. Wittig (Hrsg.): Marktplatz Internet: Von e-Learning bis e-Payment

P-73 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): “Heute schon das Morgen sehen“

P-74 Christopher Wolf, Stefan Lucks, Po-Wah Yau (Hrsg.): WEWoRC 2005 – Western European Workshop on Research in Cryptology

P-75 Jörg Desel, Ulrich Frank (Hrsg.): Enterprise Modelling and Information Systems Architecture

P-76 Thomas Kirste, Birgitta König-Riess, Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Informationssysteme – Potentiale, Hindernisse, Einsatz

P-77 Jana Dittmann (Hrsg.): SICHERHEIT 2006

P-78 K.-O. Wenkel, P. Wagner, M. Morgenstern, K. Luzi, P. Eisermann (Hrsg.): Land- und Ernährungswirtschaft im Wandel

P-79 Bettina Biel, Matthias Book, Volker Gruhn (Hrsg.): Softwareengineering 2006

P-80 Mareike Schoop, Christian Huemer, Michael Rebstock, Martin Bichler (Hrsg.): Service-Oriented Electronic Commerce

P-81 Wolfgang Karl, Jürgen Becker, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle (Hrsg.): ARCS´06

P-82 Heinrich C. Mayr, Ruth Breu (Hrsg.): Modellierung 2006

P-83 Daniel Huson, Oliver Kohlbacher, Andrei Lupas, Kay Nieselt and Andreas Zell (eds.): German Conference on Bioinformatics

P-84 Dimitris Karagiannis, Heinrich C. Mayr, (Hrsg.): Information Systems Technology and its Applications

P-85 Witold Abramowicz, Heinrich C. Mayr, (Hrsg.): Business Information Systems

P-86 Robert Krimmer (Ed.): Electronic Voting 2006

P-87 Max Mühlhäuser, Guido Rößling, Ralf Steinmetz (Hrsg.): DELFI 2006: 4. e-Learning Fachtagung Informatik

P-88 Robert Hirschfeld, Andreas Polze, Ryszard Kowalczyk (Hrsg.): NODe 2006, GSEM 2006

P-90 Joachim Schelp, Robert Winter, Ulrich Frank, Bodo Rieger, Klaus Turowski (Hrsg.): Integration, Informationslogistik und Architektur

P-91 Henrik Stormer, Andreas Meier, Michael Schumacher (Eds.): European Conference on eHealth 2006

P-92 Fernand Feltz, Benoît Otjacques, Andreas Oberweis, Nicolas Poussing (Eds.): AIM 2006

P-93 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 1

P-94 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 2

P-95 Matthias Weske, Markus Nüttgens (Eds.): EMISA 2005: Methoden, Konzepte und Technologien für die Entwicklung von dienstbasierten Informationssystemen

P-96 Saartje Brockmans, Jürgen Jung, York Sure (Eds.): Meta-Modelling and Ontologies

P-97 Oliver Göbel, Dirk Schadt, Sandra Frings, Hardo Hase, Detlef Günther, Jens Nedon (Eds.): IT-Incident Mangament & IT-Forensics – IMF 2006

P-98 Hans Brandt-Pook, Werner Simonsmeier und Thorsten Spitta (Hrsg.): Beratung in der Softwareentwicklung – Modelle, Methoden, Best Practices

P-99 Andreas Schwill, Carsten Schulte, Marco Thomas (Hrsg.): Didaktik der Informatik

P-100 Peter Forbrig, Günter Siegel, Markus Schneider (Hrsg.): HDI 2006: Hochschuldidaktik der Informatik

P-101 Stefan Böttinger, Ludwig Theuvsen, Susanne Rank, Marlies Morgenstern (Hrsg.): Agrarinformatik im Spannungsfeld zwischen Regionalisierung und globalen Wertschöpfungsketten

P-102 Otto Spaniol (Eds.): Mobile Services and Personalized Environments

P-103 Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, Christoph Brochhaus (Hrsg.): Datenbanksysteme in Business, Technologie und Web (BTW 2007)

P-104 Birgitta König-Ries, Franz Lehner, Rainer Malaka, Can Türker (Hrsg.) MMS 2007: Mobilität und mobile Informationssysteme

P-105 Wolf-Gideon Bleek, Jörg Raasch, Heinz Züllighoven (Hrsg.) Software Engineering 2007

P-106 Wolf-Gideon Bleek, Henning Schwentner, Heinz Züllighoven (Hrsg.) Software Engineering 2007 – Beiträge zu den Workshops

P-107 Heinrich C. Mayr, Dimitris Karagiannis (eds.) Information Systems Technology and its Applications

P-108 Arslan Brömme, Christoph Busch, Detlef Hühnlein (eds.) BIOSIG 2007: Biometrics and Electronic Signatures

P-109 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.) INFORMATIK 2007 Informatik trifft Logistik Band 1

P-110 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.) INFORMATIK 2007 Informatik trifft Logistik Band 2

P-111 Christian Eibl, Johannes Magenheim, Sigrid Schubert, Martin Wessner (Hrsg.) DeLFI 2007: 5. e-Learning Fachtagung Informatik

P-112 Sigrid Schubert (Hrsg.) Didaktik der Informatik in Theorie und Praxis

P-113 Sören Auer, Christian Bizer, Claudia Müller, Anna V. Zhdanova (Eds.) The Social Semantic Web 2007 Proceedings of the 1st Conference on Social Semantic Web (CSSW)

P-114 Sandra Frings, Oliver Göbel, Detlef Günther, Hardo G. Hase, Jens Nedon, Dirk Schadt, Arslan Brömme (Eds.) IMF2007 IT-incident management & IT-forensics Proceedings of the 3rd International Conference on IT-Incident Management & IT-Forensics

P-115 Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron and Dirk Walther (Eds.) German conference on bioinformatics GCB 2007

P-116 Witold Abramowicz, Leszek Maciszek (Eds.) Business Process and Services Computing 1st International Working Conference on Business Process and Services Computing BPSC 2007

P-117 Ryszard Kowalczyk (Ed.) Grid service engineering and manegement The 4th International Conference on Grid Service Engineering and Management GSEM 2007

P-118 Andreas Hein, Wilfried Thoben, Hans-Jürgen Appelrath, Peter Jensch (Eds.) European Conference on ehealth 2007

P-119 Manfred Reichert, Stefan Strecker, Klaus Turowski (Eds.) Enterprise Modelling and Information Systems Architectures Concepts and Applications

P-120 Adam Pawlak, Kurt Sandkuhl, Wojciech Cholewa, Leandro Soares Indrusiak (Eds.) Coordination of Collaborative Engineering - State of the Art and Future Challenges

P-121 Korbinian Herrmann, Bernd Bruegge (Hrsg.) Software Engineering 2008 Fachtagung des GI-Fachbereichs Softwaretechnik

P-122 Walid Maalej, Bernd Bruegge (Hrsg.) Software Engineering 2008 - Workshopband Fachtagung des GI-Fachbereichs Softwaretechnik

P-123 Michael H. Breitner, Martin Breunig, Elgar Fleisch, Ley Pousttchi, Klaus Turowski (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Technologien, Prozesse, Marktfähigkeit
Proceedings zur 3. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2008)

P-124 Wolfgang E. Nagel, Rolf Hoffmann, Andreas Koch (Eds.)
9th Workshop on Parallel Systems and Algorithms (PASA)
Workshop of the GI/ITG Speciel Interest Groups PARS and PARVA

P-125 Rolf A.E. Müller, Hans-H. Sundermeier, Ludwig Theuvsen, Stephanie Schütze, Marlies Morgenstern (Hrsg.)
Unternehmens-IT:
Führungsinstrument oder Verwaltungsbürde
Referate der 28. GIL Jahrestagung

P-126 Rainer Gimnich, Uwe Kaiser, Jochen Quante, Andreas Winter (Hrsg.)
10th Workshop Software Reengineering (WSR 2008)

P-127 Thomas Kühne, Wolfgang Reisig, Friedrich Steimann (Hrsg.)
Modellierung 2008

P-128 Ammar Alkassar, Jörg Siekmann (Hrsg.)
Sicherheit 2008
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
2.-4. April 2008
Saarbrücken, Germany

P-129 Wolfgang Hesse, Andreas Oberweis (Eds.)
Sigsand-Europe 2008
Proceedings of the Third AIS SIGSAND European Symposium on Analysis, Design, Use and Societal Impact of Information Systems

P-130 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
1. DFN-Forum Kommunikations-technologien Beiträge der Fachtagung

P-131 Robert Krimmer, Rüdiger Grimm (Eds.)
3rd International Conference on Electronic Voting 2008
Co-organized by Council of Europe, Gesellschaft für Informatik and E-Voting. CC

P-132 Silke Seehusen, Ulrike Lucke, Stefan Fischer (Hrsg.)
DeLFI 2008:
Die 6. e-Learning Fachtagung Informatik

P-133 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 1

P-134 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 2

P-135 Torsten Brinda, Michael Fothe, Peter Hubwieser, Kirsten Schlüter (Hrsg.)
Didaktik der Informatik –
Aktuelle Forschungsergebnisse

P-136 Andreas Beyer, Michael Schroeder (Eds.)
German Conference on Bioinformatics GCB 2008

P-137 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2008: Biometrics and Electronic Signatures

P-138 Barbara Dinter, Robert Winter, Peter Chamoni, Norbert Gronau, Klaus Turowski (Hrsg.)
Synergien durch Integration und Informationslogistik
Proceedings zur DW2008

P-139 Georg Herzwurm, Martin Mikusz (Hrsg.)
Industrialisierung des Software-Managements
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschaftsinformatik

P-140 Oliver Göbel, Sandra Frings, Detlef Günther, Jens Nedon, Dirk Schadt (Eds.)
IMF 2008 - IT Incident Management & IT Forensics

P-141 Peter Loos, Markus Nüttgens, Klaus Turowski, Dirk Werth (Hrsg.)
Modellierung betrieblicher Informations-systeme (MobIS 2008)
Modellierung zwischen SOA und Compliance Management

P-142 R. Bill, P. Korduan, L. Theuvsen, M. Morgenstern (Hrsg.)
Anforderungen an die Agrarinformatik durch Globalisierung und Klimaveränderung

P-143 Peter Liggesmeyer, Gregor Engels, Jürgen Münch, Jörg Dörr, Norman Riegel (Hrsg.)
Software Engineering 2009
Fachtagung des GI-Fachbereichs Softwaretechnik

P-144 Johann-Christoph Freytag, Thomas Ruf, Wolfgang Lehner, Gottfried Vossen (Hrsg.)
Datenbanksysteme in Business, Technologie und Web (BTW)

P-145 Knut Hinkelmann, Holger Wache (Eds.)
WM2009: 5th Conference on Professional Knowledge Management

P-146 Markus Bick, Martin Breunig, Hagen Höpfner (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Entwicklung, Implementierung und Anwendung
4. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2009)

P-147 Witold Abramowicz, Leszek Maciaszek, Ryszard Kowalczyk, Andreas Speck (Eds.)
Business Process, Services Computing and Intelligent Service Management
BPSC 2009 · ISM 2009 · YRW-MBP 2009

P-148 Christian Erfurth, Gerald Eichler, Volkmar Schau (Eds.)
9th International Conference on Innovative Internet Community Systems
I^{2}CS 2009

P-149 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
2. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-150 Jürgen Münch, Peter Liggesmeyer (Hrsg.)
Software Engineering
2009 - Workshopband

P-151 Armin Heinzl, Peter Dadam, Stefan Kirn, Peter Lockemann (Eds.)
PRIMIUM
Process Innovation for Enterprise Software

P-152 Jan Mendling, Stefanie Rinderle-Ma, Werner Esswein (Eds.)
Enterprise Modelling and Information Systems Architectures
Proceedings of the 3rd Int'l Workshop EMISA 2009

P-153 Andreas Schwill, Nicolas Apostolopoulos (Hrsg.)
Lernen im Digitalen Zeitalter
DeLFI 2009 – Die 7. E-Learning Fachtagung Informatik

P-154 Stefan Fischer, Erik Maehle Rüdiger Reischuk (Hrsg.)
INFORMATIK 2009
Im Focus das Leben

P-155 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2009:
Biometrics and Electronic Signatures
Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-156 Bernhard Koerber (Hrsg.)
Zukunft braucht Herkunft
25 Jahre »INFOS – Informatik und Schule«

P-157 Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, Peter Stadler (Eds.)
German Conference on Bioinformatics 2009

P-158 W. Claupein, L. Theuvsen, A. Kämpf, M. Morgenstern (Hrsg.)
Precision Agriculture Reloaded – Informationsgestützte Landwirtschaft

P-159 Gregor Engels, Markus Luckey, Wilhelm Schäfer (Hrsg.)
Software Engineering 2010

P-160 Gregor Engels, Markus Luckey, Alexander Pretschner, Ralf Reussner (Hrsg.)
Software Engineering 2010 – Workshopband
(inkl. Doktorandensymposium)

P-161 Gregor Engels, Dimitris Karagiannis Heinrich C. Mayr (Hrsg.)
Modellierung 2010

P-162 Maria A. Wimmer, Uwe Brinkhoff, Siegfried Kaiser, Dagmar Lück-Schneider, Erich Schweighofer, Andreas Wiebe (Hrsg.)
Vernetzte IT für einen effektiven Staat
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010

P-163 Markus Bick, Stefan Eulgem, Elgar Fleisch, J. Felix Hampe, Birgitta König-Ries, Franz Lehner, Key Pousttchi, Kai Rannenberg (Hrsg.)
Mobile und Ubiquitäre Informationssysteme
Technologien, Anwendungen und Dienste zur Unterstützung von mobiler Kollaboration

P-164 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2010: Biometrics and Electronic Signatures Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-165 Gerald Eichler, Peter Kropf, Ulrike Lechner, Phayung Meesad, Herwig Unger (Eds.)
10th International Conference on Innovative Internet Community Systems (I^2CS) – Jubilee Edition 2010 –

P-166 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
3. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-167 Robert Krimmer, Rüdiger Grimm (Eds.)
4th International Conference on Electronic Voting 2010
co-organized by the Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-168 Ira Diethelm, Christina Dörge, Claudia Hildebrandt, Carsten Schulte (Hrsg.)
Didaktik der Informatik
Möglichkeiten empirischer Forschungsmethoden und Perspektiven der Fachdidaktik

P-169 Michael Kerres, Nadine Ojstersek Ulrik Schroeder, Ulrich Hoppe (Hrsg.)
DeLFI 2010 - 8. Tagung der Fachgruppe E-Learning der Gesellschaft für Informatik e.V.

P-170 Felix C. Freiling (Hrsg.)
Sicherheit 2010
Sicherheit, Schutz und Zuverlässigkeit

P-171 Werner Esswein, Klaus Turowski, Martin Juhrisch (Hrsg.)
Modellierung betrieblicher Informationssysteme (MobIS 2010)
Modellgestütztes Management

P-172 Stefan Klink, Agnes Koschmider Marco Mevius, Andreas Oberweis (Hrsg.)
EMISA 2010
Einflussfaktoren auf die Entwicklung flexibler, integrierter Informationssysteme
Beiträge des Workshops der GI-Fachgruppe EMISA (Entwicklungsmethoden für Informationssysteme und deren Anwendung)

P-173 Dietmar Schomburg, Andreas Grote (Eds.)
German Conference on Bioinformatics 2010

P-174 Arslan Brömme, Torsten Eymann, Detlef Hühnlein, Heiko Roßnagel, Paul Schmücker (Hrsg.)
perspeGKtive 2010
Workshop „Innovative und sichere Informationstechnologie für das Gesundheitswesen von morgen“

P-175 Klaus-Peter Fähnrich, Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für die Informatik
Band 1

P-176 Klaus-Peter Fähnrich, Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für die Informatik
Band 2

P-177 Witold Abramowicz, Rainer Alt, Klaus-Peter Fähnrich, Bogdan Franczyk, Leszek A. Maciaszek (Eds.)
INFORMATIK 2010
Business Process and Service Science – Proceedings of ISSS and BPSC

P-178 Wolfram Pietsch, Benedikt Krams (Hrsg.)
Vom Projekt zum Produkt
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschafts-informatik (WI-MAW), Aachen, 2010

P-179 Stefan Gruner, Bernhard Rumpe (Eds.)
FM+AM`2010
Second International Workshop on Formal Methods and Agile Methods

P-180 Theo Härder, Wolfgang Lehner, Bernhard Mitschang, Harald Schöning, Holger Schwarz (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW)
14. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme“ (DBIS)

P-181 Michael Clasen, Otto Schätzel, Brigitte Theuvsen (Hrsg.)
Qualität und Effizienz durch informationsgestützte Landwirtschaft, Fokus: Moderne Weinwirtschaft

P-182 Ronald Maier (Hrsg.)
6th Conference on Professional Knowledge Management
From Knowledge to Action

P-183 Ralf Reussner, Matthias Grund, Andreas Oberweis, Walter Tichy (Hrsg.)
Software Engineering 2011
Fachtagung des GI-Fachbereichs Softwaretechnik

P-184 Ralf Reussner, Alexander Pretschner, Stefan Jähnichen (Hrsg.)
Software Engineering 2011
Workshopband
(inkl. Doktorandensymposium)

P-185 Hagen Höpfner, Günther Specht, Thomas Ritz, Christian Bunse (Hrsg.)
MMS 2011: Mobile und ubiquitäre Informationssysteme Proceedings zur 6. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2011)

P-186 Gerald Eichler, Axel Küpper, Volkmar Schau, Hacène Fouchal, Herwig Unger (Eds.)
11th International Conference on Innovative Internet Community Systems (I^2CS)

P-187 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
4. DFN-Forum Kommunikations-technologien, Beiträge der Fachtagung 20. Juni bis 21. Juni 2011 Bonn

P-188 Holger Rohland, Andrea Kienle, Steffen Friedrich (Hrsg.)
DeLFI 2011 – Die 9. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. 5.–8. September 2011, Dresden

P-189 Thomas, Marco (Hrsg.)
Informatik in Bildung und Beruf
INFOS 2011
14. GI-Fachtagung Informatik und Schule

P-190 Markus Nüttgens, Oliver Thomas, Barbara Weber (Eds.)
Enterprise Modelling and Information Systems Architectures (EMISA 2011)

P-191 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2011
International Conference of the Biometrics Special Interest Group

P-192 Hans-Ulrich Heiß, Peter Pepper, Holger Schlingloff, Jörg Schneider (Hrsg.)
INFORMATIK 2011
Informatik schafft Communities

P-193 Wolfgang Lehner, Gunther Piller (Hrsg.)
IMDM 2011

P-194 M. Clasen, G. Fröhlich, H. Bernhardt, K. Hildebrand, B. Theuvsen (Hrsg.)
Informationstechnologie für eine nachhaltige Landbewirtschaftung
Fokus Forstwirtschaft

P-195 Neeraj Suri, Michael Waidner (Hrsg.)
Sicherheit 2012
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 6. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)

P-196 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2012
Proceedings of the 11th International Conference of the Biometrics Special Interest Group

P-197 Jörn von Lucke, Christian P. Geiger, Siegfried Kaiser, Erich Schweighofer, Maria A. Wimmer (Hrsg.)
Auf dem Weg zu einer offenen, smarten und vernetzten Verwaltungskultur
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2012

P-198 Stefan Jähnichen, Axel Küpper, Sahin Albayrak (Hrsg.)
Software Engineering 2012
Fachtagung des GI-Fachbereichs Softwaretechnik

P-199 Stefan Jähnichen, Bernhard Rumpe, Holger Schlingloff (Hrsg.)
Software Engineering 2012
Workshopband

P-200 Gero Mühl, Jan Richling, Andreas Herkersdorf (Hrsg.)
ARCS 2012 Workshops

P-201 Elmar J. Sinz Andy Schürr (Hrsg.)
Modellierung 2012

P-202 Andrea Back, Markus Bick, Martin Breunig, Key Pousttchi, Frédéric Thiesse (Hrsg.)
MMS 2012:Mobile und Ubiquitäre Informationssysteme

P-203 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
5. DFN-Forum Kommunikations-technologien
Beiträge der Fachtagung

P-204 Gerald Eichler, Leendert W. M. Wienhofen, Anders Kofod-Petersen, Herwig Unger (Eds.)
12th International Conference on Innovative Internet Community Systems (I2CS 2012)

P-205 Manuel J. Kripp, Melanie Volkamer, Rüdiger Grimm (Eds.)
5th International Conference on Electronic Voting 2012 (EVOTE2012)
Co-organized by the Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-206 Stefanie Rinderle-Ma, Mathias Weske (Hrsg.)
EMISA 2012
Der Mensch im Zentrum der Modellierung

P-207 Jörg Desel, Jörg M. Haake, Christian Spannagel (Hrsg.)
DeLFI 2012: Die 10. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V.
24.–26. September 2012

P-208 Ursula Goltz, Marcus Magnor, Hans-Jürgen Appelrath, Herbert Matthies, Wolf-Tilo Balke, Lars Wolf (Hrsg.) INFORMATIK 2012

P-209 Hans Brandt-Pook, André Fleer, Thorsten Spitta, Malte Wattenberg (Hrsg.) Nachhaltiges Software Management

P-210 Erhard Plödereder, Peter Dencker, Herbert Klenk, Hubert B. Keller, Silke Spitzer (Hrsg.) Automotive – Safety & Security 2012 Sicherheit und Zuverlässigkeit für automobile Informationstechnik

P-211 M. Clasen, K. C. Kersebaum, A. Meyer-Aurich, B. Theuvsen (Hrsg.) Massendatenmanagement in der Agrar- und Ernährungswirtschaft Erhebung - Verarbeitung - Nutzung Referate der 33. GIL-Jahrestagung 20. – 21. Februar 2013, Potsdam

P-212 Arslan Brömme, Christoph Busch (Eds.) BIOSIG 2013 Proceedings of the 12th International Conference of the Biometrics Special Interest Group 04.–06. September 2013 Darmstadt, Germany

P-213 Stefan Kowalewski, Bernhard Rumpe (Hrsg.) Software Engineering 2013 Fachtagung des GI-Fachbereichs Softwaretechnik

P-214 Volker Markl, Gunter Saake, Kai-Uwe Sattler, Gregor Hackenbroich, Bernhard Mit schang, Theo Härder, Veit Köppen (Hrsg.) Datenbanksysteme für Business, Technologie und Web (BTW) 2013 13. – 15. März 2013, Magdeburg

P-215 Stefan Wagner, Horst Lichter (Hrsg.) Software Engineering 2013 Workshopband (inkl. Doktorandensymposium) 26. Februar – 1. März 2013, Aachen

P-216 Gunter Saake, Andreas Henrich, Wolfgang Lehner, Thomas Neumann, Veit Köppen (Hrsg.) Datenbanksysteme für Business, Technologie und Web (BTW) 2013 – Workshopband 11. – 12. März 2013, Magdeburg

P-217 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.) 6. DFN-Forum Kommunikations-technologien Beiträge der Fachtagung 03.–04. Juni 2013, Erlangen

P-218 Andreas Breiter, Christoph Rensing (Hrsg.) DeLFI 2013: Die 11 e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. (GI) 8. – 11. September 2013, Bremen

P-219 Norbert Breier, Peer Stechert, Thomas Wilke (Hrsg.) Informatik erweitert Horizonte INFOS 2013 15. GI-Fachtagung Informatik und Schule 26. – 28. September 2013

P-220 Matthias Horbach (Hrsg.) INFORMATIK 2013 Informatik angepasst an Mensch, Organisation und Umwelt 16. – 20. September 2013, Koblenz

P-221 Maria A. Wimmer, Marijn Janssen, Ann Macintosh, Hans Jochen Scholl, Efthimios Tambouris (Eds.) Electronic Government and Electronic Participation Joint Proceedings of Ongoing Research of IFIP EGOV and IFIP ePart 2013 16. – 19. September 2013, Koblenz

P-222 Reinhard Jung, Manfred Reichert (Eds.) Enterprise Modelling and Information Systems Architectures (EMISA 2013) St. Gallen, Switzerland September 5. – 6. 2013

P-223 Detlef Hühnlein, Heiko Roßnagel (Hrsg.) Open Identity Summit 2013 10. – 11. September 2013 Kloster Banz, Germany

P-224 Eckhart Hanser, Martin Mikusz, Masud Fazal-Baqaie (Hrsg.) Vorgehensmodelle 2013 Vorgehensmodelle – Anspruch und Wirklichkeit 20. Tagung der Fachgruppe Vorgehensmodelle im Fachgebiet Wirtschaftsinformatik (WI-VM) der Gesellschaft für Informatik e.V. Lörrach, 2013

P-225 Hans-Georg Fill, Dimitris Karagiannis, Ulrich Reimer (Hrsg.) Modellierung 2014 19. – 21. März 2014, Wien

P-226 M. Clasen, M. Hamer, S. Lehnert, B. Petersen, B. Theuvsen (Hrsg.) IT-Standards in der Agrar- und Ernährungswirtschaft Fokus: Risiko- und Krisenmanagement Referate der 34. GIL-Jahrestagung 24. – 25. Februar 2014, Bonn

P-227 Wilhelm Hasselbring,
Nils Christian Ehmke (Hrsg.)
Software Engineering 2014
Fachtagung des GI-Fachbereichs Softwaretechnik
25. – 28. Februar 2014
Kiel, Deutschland

P-228 Stefan Katzenbeisser, Volkmar Lotz, Edgar Weippl (Hrsg.)
Sicherheit 2014
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 7. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
19. – 21. März 2014, Wien

P-229 Dagmar Lück-Schneider, Thomas Gordon, Siegfried Kaiser, Jörn von Lucke,Erich Schweighofer, Maria A.Wimmer, Martin G. Löhe (Hrsg.)
Gemeinsam Electronic Government ziel(gruppen)gerecht gestalten und organisieren
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI)
2014, 20.-21. März 2014 in Berlin

P-230 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2014
Proceedings of the 13th International Conference of the Biometrics Special Interest Group
10. – 12. September 2014 in Darmstadt, Germany

P-231 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
7. DFN-Forum Kommunikationstechnologien
16. – 17. Juni 2014
Fulda

P-232 E. Plödereder, L. Grunske, E. Schneider, D. Ull (Hrsg.)
INFORMATIK 2014
Big Data – Komplexität meistern
22. – 26. September 2014
Stuttgart

P-233 Stephan Trahasch, Rolf Plötzner, Gerhard Schneider, Claudia Gayer, Daniel Sassiat, Nicole Wöhrle (Hrsg.)
DeLFI 2014 – Die 12. e-Learning Fachtagung Informatik
der Gesellschaft für Informatik e.V.
15. – 17. September 2014
Freiburg

P-234 Fernand Feltz, Bela Mutschler, Benoît Otjacques (Eds.)
Enterprise Modelling and Information Systems Architectures
(EMISA 2014)
Luxembourg, September 25-26, 2014

P-235 Robert Giegerich, Ralf Hofestädt, Tim W. Nattkemper (Eds.)
German Conference on Bioinformatics 2014
September 28 – October 1
Bielefeld, Germany

P-236 Martin Engstler, Eckhart Hanser, Martin Mikusz, Georg Herzwurm (Hrsg.)
Projektmanagement und Vorgehensmodelle 2014
Soziale Aspekte und Standardisierung
Gemeinsame Tagung der Fachgruppen Projektmanagement (WI-PM) und Vorgehensmodelle (WI-VM) im Fachgebiet Wirtschaftsinformatik der Gesellschaft für Informatik e.V., Stuttgart 2014

P-237 Detlef Hühnlein, Heiko Roßnagel (Hrsg.)
Open Identity Summit 2014
4.–6. November 2014
Stuttgart, Germany

P-238 Arno Ruckelshausen, Hans-Peter Schwarz, Brigitte Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Referate der 35. GIL-Jahrestagung
23. – 24. Februar 2015, Geisenheim

P-239 Uwe Aßmann, Birgit Demuth, Thorsten Spitta, Georg Püschel, Ronny Kaiser (Hrsg.)
Software Engineering & Management 2015
17.-20. März 2015, Dresden

P-240 Herbert Klenk, Hubert B. Keller, Erhard Plödereder, Peter Dencker (Hrsg.)
Automotive – Safety & Security 2015
Sicherheit und Zuverlässigkeit für automobile Informationstechnik
21.–22. April 2015, Stuttgart

P-241 Thomas Seidl, Norbert Ritter, Harald Schöning, Kai-Uwe Sattler, Theo Härder, Steffen Friedrich, Wolfram Wingerath (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2015)
04. – 06. März 2015, Hamburg

P-242 Norbert Ritter, Andreas Henrich, Wolfgang Lehner, Andreas Thor, Steffen Friedrich, Wolfram Wingerath (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2015) – Workshopband
02. – 03. März 2015, Hamburg

P-243 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
8. DFN-Forum Kommunikationstechnologien
06.–09. Juni 2015, Lübeck

P-244 Alfred Zimmermann, Alexander Rossmann (Eds.)
Digital Enterprise Computing (DEC 2015)
Böblingen, Germany June 25-26, 2015

P-245 Arslan Brömme, Christoph Busch , Christian Rathgeb, Andreas Uhl (Eds.)
BIOSIG 2015
Proceedings of the 14th International Conference of the Biometrics Special Interest Group
09.–11. September 2015
Darmstadt, Germany

P-246 Douglas W. Cunningham, Petra Hofstedt, Klaus Meer, Ingo Schmitt (Hrsg.)
INFORMATIK 2015
28.9.-2.10. 2015, Cottbus

P-247 Hans Pongratz, Reinhard Keil (Hrsg.)
DeLFI 2015 – Die 13. E-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. (GI)
1.–4. September 2015
München

P-248 Jens Kolb, Henrik Leopold, Jan Mendling (Eds.)
Enterprise Modelling and Information Systems Architectures
Proceedings of the 6th Int. Workshop on Enterprise Modelling and Information Systems Architectures, Innsbruck, Austria
September 3-4, 2015

P-249 Jens Gallenbacher (Hrsg.)
Informatik allgemeinbildend begreifen
INFOS 2015 16. GI-Fachtagung Informatik und Schule
20.–23. September 2015

P-250 Martin Engstler, Masud Fazal-Baqaie, Eckhart Hanser, Martin Mikusz, Alexander Volland (Hrsg.)
Projektmanagement und Vorgehensmodelle 2015
Hybride Projektstrukturen erfolgreich umsetzen
Gemeinsame Tagung der Fachgruppen Projektmanagement (WI-PM) und Vorgehensmodelle (WI-VM) im Fachgebiet Wirtschaftsinformatik der Gesellschaft für Informatik e.V., Elmshorn 2015

P-251 Detlef Hühnlein, Heiko Roßnagel, Raik Kuhlisch, Jan Ziesing (Eds.)
Open Identity Summit 2015
10.–11. November 2015
Berlin, Germany

P-252 Jens Knoop, Uwe Zdun (Hrsg.)
Software Engineering 2016
Fachtagung des GI-Fachbereichs Softwaretechnik
23.–26. Februar 2016, Wien

P-253 A. Ruckelshausen, A. Meyer-Aurich, T. Rath, G. Recke, B. Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Fokus: Intelligente Systeme – Stand der Technik und neue Möglichkeiten
Referate der 36. GIL-Jahrestagung
22.-23. Februar 2016, Osnabrück

P-254 Andreas Oberweis, Ralf Reussner (Hrsg.)
Modellierung 2016
2.–4. März 2016, Karlsruhe

P-255 Stefanie Betz, Ulrich Reimer (Hrsg.)
Modellierung 2016 Workshopband
2.–4. März 2016, Karlsruhe

P-256 Michael Meier, Delphine Reinhardt, Steffen Wendzel (Hrsg.)
Sicherheit 2016
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 8. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
5.–7. April 2016, Bonn

P-257 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
9. DFN-Forum Kommunikationstechnologien
31. Mai – 01. Juni 2016, Rostock

P-258 Dieter Hertweck, Christian Decker (Eds.)
Digital Enterprise Computing (DEC 2016)
14.–15. Juni 2016, Böblingen

P-259 Heinrich C. Mayr, Martin Pinzger (Hrsg.)
INFORMATIK 2016
26.–30. September 2016, Klagenfurt

P-260 Arslan Brömme, Christoph Busch, Christian Rathgeb, Andreas Uhl (Eds.)
BIOSIG 2016
Proceedings of the 15th International Conference of the Biometrics Special Interest Group
21.–23. September 2016, Darmstadt

P-261 Detlef Rätz, Michael Breidung, Dagmar Lück-Schneider, Siegfried Kaiser, Erich Schweighofer (Hrsg.)
Digitale Transformation: Methoden, Kompetenzen und Technologien für die Verwaltung
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2016
22.–23. September 2016, Dresden

P-262 Ulrike Lucke, Andreas Schwill, Raphael Zender (Hrsg.)
DeLFI 2016 – Die 14. E-Learning Fachtagung Informatik
der Gesellschaft für Informatik e.V. (GI)
11.–14. September 2016, Potsdam

P-263 Martin Engstler, Masud Fazal-Baqaie, Eckhart Hanser, Oliver Linssen, Martin Mikusz, Alexander Volland (Hrsg.)
Projektmanagement und Vorgehensmodelle 2016
Arbeiten in hybriden Projekten: Das Sowohl-als-auch von Stabilität und Dynamik
Gemeinsame Tagung der Fachgruppen Projektmanagement (WI-PM) und Vorgehensmodelle (WI-VM) im Fachgebiet Wirtschaftsinformatik
der Gesellschaft für Informatik e.V., Paderborn 2016

P-264 Detlef Hühnlein, Heiko Roßnagel, Christian H. Schunck, Maurizio Talamo (Eds.)
Open Identity Summit 2016
der Gesellschaft für Informatik e.V. (GI)
13.–14. October 2016, Rome, Italy

P-265 Bernhard Mitschang, Daniela Nicklas,Frank Leymann, Harald Schöning, Melanie Herschel, Jens Teubner, Theo Härder, Oliver Kopp, Matthias Wieland (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2017)
6.–10. März 2017, Stuttgart

P-266 Bernhard Mitschang, Norbert Ritter, Holger Schwarz, Meike Klettke, Andreas Thor, Oliver Kopp, Matthias Wieland (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW 2017)
Workshopband
6.–7. März 2017, Stuttgart

P-267 Jan Jürjens, Kurt Schneider (Hrsg.)
Software Engineering 2017
21.–24. Februar 2017, Hannover

P-268 A. Ruckelshausen, A. Meyer-Aurich, W. Lentz, B. Theuvsen (Hrsg.)
Informatik in der Land-, Forst- und Ernährungswirtschaft
Fokus: Digitale Transformation – Wege in eine zukunftsfähige Landwirtschaft
Referate der 37. GIL-Jahrestagung
06.–07. März 2017, Dresden

P-269 Peter Dencker, Herbert Klenk, Hubert Keller, Erhard Plödereder (Hrsg.)
Automotive – Safety & Security 2017
30.–31. Mai 2017, Stuttgart

P-270 Arslan Brömme, Christoph Busch, Antitza Dantcheva, Christian Rathgeb, Andreas Uhl (Eds.)
BIOSIG 2017
20.–22. September 2017, Darmstadt

P-271 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
10. DFN-Forum Kommunikations-technologien
30. – 31. Mai 2017, Berlin

P-272 Alexander Rossmann, Alfred Zimmermann (eds.)
Digital Enterprise Computing (DEC 2017)
11.–12. Juli 2017, Böblingen

P-273 Christoph Igel, Carsten Ullrich,
Martin Wessner (Hrsg.)
BILDUNGSRÄUME
DeLFI 2017
Die 15. e-Learning Fachtagung Informatik
der Gesellschaft für Informatik e.V. (GI)
5. bis 8. September 2017, Chemnitz

P-274 Ira Diethelm (Hrsg.)
Informatische Bildung zum Verstehen
und Gestalten der digitalen Welt
13.–15. September 2017, Oldenburg

P-275 Maximilian Eibl, Martin Gaedke (Hrsg.)
INFORMATIK 2017
25.–29. September 2017, Chemnitz

P-277 Lothar Fritsch, Heiko Roßnagel,
Detlef Hühnlein (Hrsg.)
Open Identity Summit 2017
5.– 6. October 2017, Karlstad, Sweden

The titles can be purchased at:

Köllen Druck + Verlag GmbH
Ernst-Robert-Curtius-Str. 14 · D-53117 Bonn
Fax: +49 (0)228/9898222
E-Mail: druckverlag@koellen.de