

Proceedings of the First European Workshop on
Biometrics and
Identity Management
(BIOID 2008)

Ben Schouten
Niels Christian Juul
(Editors)



Copyright © 2008

Ben Schouten and Niels Christian Juul



Computer Science
Roskilde University
P. O. Box 260
DK-4000 Roskilde
Denmark

Telephone: +45 4674 3839

Telefax: +45 4674 3072

Internet: http://www.ruc.dk/dat_en/

E-mail: datalogi@ruc.dk

All rights reserved

Permission to copy, print, or redistribute all or part of this work is granted for educational or research use on condition that this copyright notice is included in any copy.

ISSN 0109-9779

Research reports are available electronically from:

http://www.ruc.dk/dat_en/research/reports/

Ben Schouten, Niels Christian Juul (Eds.)

Biometrics and Identity Management

First European Workshop on Biometrics and Identity Management,
BIOID 2008
Roskilde University, Denmark, 7-9 May 2008
Revised Selected Papers



Printed at Roskilde University, Denmark

Preface

A key driving factor for biometrics is the widespread national and international deployment of biometric systems that has been initiated in the past two years and is about to accelerate. While nearly all current biometric deployments are government-led and principally concerned with national security and border control scenarios it is now apparent that the widespread availability of biometrics in everyday life will also spin out an ever increasing number of (private) applications in other domains. Crucial to this vision is the management of the user's identity, which does not only imply the creation and update of a biometric template, but requires the development of instruments to properly handle all the data and operations related to the user identity.

These proceedings contain the selected and revised papers that were presented during the first European Workshop on Biometrics and Identity management. BIOID 2008. The papers are categorized in four classes. These classes represent the 4 working groups of the COST Action 2101. For more information, see <http://cost2101.org/>

1. Biometric data quality and multimodal biometric templates,
2. Unsupervised interactive interfaces for multimodal biometrics,
3. Biometric attacks and countermeasures,
4. Standards and privacy issues for biometrics in identity documents and smart cards.

BIOID 2008 is an initiative of the COST Action 2101 on Biometrics for Identity Documents and Smart Cards. It is supported by the EU Framework 7 Programme. Other sponsors of the Workshop are: The European Biometrics Forum, The Danish Biometrics Research Project Consortium, the UK Biometrics Institute and the Institution of Engineering and Technology.

The BIOID workshop was jointly organized and held at the Roskilde University in Denmark from May 7- May 9, 2008.

May 2008

Ben Schouten,
Andrzej Drygajlo,
Niels Christian Juul and
Michael Fairhurst

Organization

Chairs

Ben Schouten	CWI/Fontys, NL
Andrzej Drygajlo	EPFL, CH
Niels Christian Juul	Roskilde University, DK
Michael Fairhurst	University of Kent, UK

Workshop Organization

Niels Christian Juul (Chair)	Roskilde University, DK
Agnete Nebsager	Roskilde University, DK
Heidi Lundquist	Roskilde University, DK

Program Committee

Ben Schouten (Chair)	CWI/Fontys, NL
Aladdin Ariyaeinia	University of Hertfordshire, UK
Nicolas Delvaux	Sagem, FR
Andrzej Drygajlo	EPFL, CH
Michael Fairhurst	University of Kent, UK
Niels Christian Juul	Roskilde University, DK
Slobodan Ribaric	University of Zagreb, HR
Albert Ali Salah	CWI, NL
Bülent Sankur	Bogazici University, TR

Sponsoring Institutions

- COST Action 2101: Biometrics for Identity Documents and Smart cards,
- European Biometrics Forum (EBF),
- The Danish Biometrics Research Project Consortium,
- UK Biometrics Institute,
- Institution of Engineering and Technology (IET),
- Fontys,
- Roskilde University

Table of Contents

Biometric Data Quality

Quality-based Score Normalization and Frame Selection for Video-based Person Authentication <i>Enrique Argones Rua, Jose Luis Alba Castro and Carmen Garcia Mateo</i>	1
Face Quality Assessment System in Video Sequences <i>Kamal Nasrollahi and Thomas B. Moeslund</i>	11
On quality of quality measures for classification <i>Krzysztof Kryszczuk and Andrzej Drygajlo</i>	21
Definition of Fingerprint Scanner Image Quality Specifications by Operational Quality <i>A. Alessandrini, R. Cappelli, M. Ferrara and D. Maltoni</i>	31

Biometrical Templates: Face Recognition

Modeling Marginal Distributions of Gabor Coefficients: Application to Biometric Template Reduction <i>Daniel González-Jiménez and José Luis Alba-Castro</i>	41
Bosphorus Database for 3D Face Analysis <i>Arman Savran, Nese Alyüz, Hamdi Dibeklioglu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur and Lale Akarun</i>	51
3D Face Recognition Benchmarks on the Bosphorus Database with Focus on Facial Expressions <i>Neşe Alyuz, Berk Gökberk, Hamdi Dibeklioglu, Arman Savran, Albert Ali Salah, Lale Akarun and Bülent Sankur</i>	62
Identity Management in Face Recognition Systems <i>Massimo Tistarelli and Enrico Grosso</i>	72
Discriminant Non-negative Matrix Factorization and Projected Gradients for Frontal Face Verification <i>Irene Kotsia, Stefanos Zafeiriou and Ioannis Pitas</i>	87

Biometrical Templates: Other Modalities

On the Discrimination Capabilities of Speech Cepstral Features <i>A. Malegaonkar, A. Ariyaeinia, P. Sivakumaran and S. Pillay</i>	95
Multimodal Speaker Identification based on Text and Speech <i>Panagiotis Moschonas and Constantine Kotropoulos</i>	104

A Palmprint Verification System Based on Phase Congruency Features <i>Vitomir Štruc and Nikola Pavešić</i>	114
Some unusual experiments with PCA-based palmprint and face recognition <i>Ivan Krevatin and Slobodan Ribarić</i>	124
An Empirical Comparison Of Individual Machine Learning Techniques In Signature And Fingerprint Classification <i>Márjory Abreu and Michael Fairhurst</i>	134
Promoting diversity in Gaussian mixture ensembles: an application to signature verification <i>Jonas Richiardi, Andrzej Drygajlo and Laetitia Todesk</i>	144

Biometric Attacks and Countermeasures

Advanced Studies on Reproducibility of Biometric Hashes <i>Tobias Scheidat, Claus Vielhauer and Jana Dittmann</i>	154
Additive Block Coding Schemes for Biometric Authentication with the DNA Data <i>Vladimir B. Balakirsky, Anahit R. Ghazaryan, A. J. Han Vinck</i>	164
Template Protection for On-line Signature-based Recognition Systems <i>Emanuele Maiorana, Patrizio Campisi, and Alessandro Neri</i>	174
Direct attacks using fake images in iris verification <i>Virginia Ruiz-Albacete, Pedro Tome-Gonzalez, Fernando Alonso-Fernandez, Javier Galbally, Julian Fierrez and Javier Ortega-Garcia</i>	184

Biometric Interfaces, Standards and Privacy

Evaluating systems assessing face-image compliance with ICAO/ISO standards <i>M. Ferrara, A. Franco and D. Maltoni</i>	194
Automatic Evaluation of Stroke Slope <i>Georgi Gluhchev and Ognyan Boumbarov</i>	204
Biometric system based on voice recognition using multiclassifiers <i>Mohamed Chenafa, Dan Istrate, Valeriu Vrabie and Michel Herbin</i>	208
POLYBIO: Multimodal Biometric Data Acquisition Platform and Security System <i>Anastasis Kounoudes, Nicolas Tsapatsoulis, Zenonas Theodosiou and Marios Milis</i>	218

Quality-based Score Normalization and Frame Selection for Video-based Person Authentication

Enrique Argones Rúa, José Luis Alba Castro, and Carmen García Mateo *

Signal Technologies Group
Signal Theory and Communications Department
University of Vigo, Spain
eargones, jalba, carmen@gts.tsc.uvigo.es

Abstract. This paper addresses the incorporation of quality measures to video-based person authentication. A theoretical framework to incorporate quality measures in biometric authentication is exposed. Two different quality-based score normalization techniques are derived from this theoretical framework. Furthermore, a quality-based frame selection technique and a new face image quality measure are also presented. The ability of this quality measure and the proposed quality-based score normalization techniques and quality-based frame selection technique to improve verification performance is experimentally evaluated in a video-based face verification experiment on the BANCA Database.

1 Introduction

Face verification is one of the most important and challenging biometric verification modalities. Face verification systems can be used in a wide variety of applications, including building access and web-based access to services among others, since low cost sensors such as web-cams can be used for image acquisition. However, face verification systems are sensitive to illumination changes, partial occlusions of the face, shadowing, changing background, low resolution problems, image noise, pose and appearance changes.

The influence of some of these factors can be diminished by increasing the intra-user variability registered in the user template. The incorporation of the full video sequence to the video recognition system in both enrolment and verification processes provides much more information than a reduced set of still images, enabling a statistically significant improvement in verification performance [1]. The video-based person authentication system used in this paper uses a Gaussian mixture model-universal background model (GMM-UBM) scheme [2]. Each location in the face is modeled by a GMM-UBM, which is adapted to the video frames in the user enrolment video by means of the MAP algorithm. This approach is able to encode the statistically discriminant information at each location of the face of the user.

* This work has been partially supported by Spanish Ministry of Education and Science (project PRESA TEC2005-07212), by the Xunta de Galicia (project PGIDIT05TIC32202PR)

However, mismatch in quality related factors between enrolment and test sessions can lead to degraded performance even though enough discriminant information is encoded in the user template. The use of quality measures [3] can somehow reduce the influence of quality mismatches in the test phase.

This paper provides a theoretical framework that incorporates the quality measures into the verification process. Starting from this theoretical framework, two quality-based score normalization techniques are derived. A new quality measure for frontal face images and a new quality-based frame selection technique are also proposed. The aim of this quality-based frame selection technique is to select the high quality frames in the test video sequence in order to improve verification performance. Experiments on the BANCA Database [4] will show the effectiveness of the combined use of the quality-based score normalization and quality-based frame selection techniques when compared to the baseline video-based face verification system.

The paper is outlined as follows: Section 2 describes the GMM-UBM video-based identity verification system. Section 3 describes the theoretical framework and techniques derived to incorporate the quality measures into the verification process. The quality-based frame selection is derived and described in Section 5. The proposed quality measure for frontal face images is presented in Section 6. The experimental framework for the experiments carried out in this paper is described in Section 7. Experiments to check the effectiveness of these techniques are shown in Section 8. Experimental results are discussed in Section 9, and paper is finally drawn to conclusions in Section 10.

2 Video-based Face Verification

The video-based face verification system first detects the face region using a face detector based on a cascade of boosted classifiers which use an extended set of Haar-like features [5]. Eyes localisation is performed using the same principle. Face is then rotated and scaled in order to set the eyes position in the same place for all the faces. Both face and eye localisations are only roughly estimated, and therefore localisation errors are present in the face image. Let us denote the video frames sequence where a face is detected as $\mathcal{V} = \{\mathcal{I}^{\mathcal{V},1}, \dots, \mathcal{I}^{\mathcal{V},N_{\mathcal{V}}}\}$, where $N_{\mathcal{V}}$ is the number of frames where a face is detected. Gabor jets [6] ($M = 40$ responses of Gabor filters with 5 scales and 8 orientations) are extracted at fixed points along a rectangular grid of dimensions $D = 10 \times 10$ superimposed on each normalized face image. Frame $\mathcal{I}^{\mathcal{V},k}$ is characterised by the moduli of all the extracted Gabor jets $\mathcal{I}^{\mathcal{V},i} = \{\mathcal{J}_1^{\mathcal{V},i}, \dots, \mathcal{J}_D^{\mathcal{V},i}\}$. The modulus of the k -th Gabor jet extracted from the i -th frame in \mathcal{V} is denoted as $\mathcal{J}_i^{\mathcal{V},k} = \{a_{i,1}^{\mathcal{V},k}, \dots, a_{i,M}^{\mathcal{V},k}\}$.

GMM-UBM verification paradigm is adapted to video-based verification: a 64 mixtures UBM is trained for each grid location and then it is adapted to the corresponding jets obtained from the user enrolment video by means of the MAP technique. Independence between the distributions of the jets from each node is assumed in order to avoid the curse of dimensionality in the UBM training. Gaussian mixtures are constrained to have diagonal covariance matrixes. The

verification score for the video \mathcal{V} and claimed identity u is computed as the following loglikelihood ratio [2]:

$$s_{\mathcal{V}} = \text{Log} \left(\prod_{i=1}^{N_{\mathcal{V}}} \prod_{k=1}^D \frac{f_{u,k}(\mathcal{J}_k^{\mathcal{V},i})}{f_{UBM,k}(\mathcal{J}_k^{\mathcal{V},i})} \right) \quad (1)$$

3 Incorporating Quality Measures in Biometric Verification

Starting from the theoretical framework exposed in [7], the verification scores S produced by classes $C = 0$ or $C = 1$ (false and true identity claims respectively) are actually computed from input vectors \mathbf{x} which are contaminated by a noise random process N that produce noise vectors \mathbf{n} : $\mathbf{x} = \Phi(\mathbf{v}, \mathbf{n})$, where \mathbf{v} are the clean vectors produced by the corresponding class. It is reasonable to assume that the verification scores S are influenced by the noise process. Unfortunately, the characteristics and influence of the noise process N in the scores S are unknown in general, since the noise value \mathbf{n} is not directly observable. However, some measures performed over the observable noisy signal \mathbf{x} can hopefully provide us with useful information about the noise. These measures are called in general *quality measures* [3].

We can model a quality measure as a random process Q that produces an output measurement q which is related to the noise \mathbf{n} present in the noisy vectors \mathbf{x} . In general, we can write $\mathbf{q} = \Omega(\mathbf{x}) = \Omega(\Phi(\mathbf{v}, \mathbf{n}))$. When we have a set of scores associated to known values of the quality measure Q and class C , the conditional probability density function $p(s|C, Q)$ can be estimated.

Since the quality measure Q depends only on the quality of the biometric signal, it is reasonable to assume that it is class independent: $p(q|C = 0) = p(q|C = 1)$. Besides, if we assume equiprobable classes then $P(C = 0) = P(C = 1)$, and thus the Bayesian verification decision is taken by:

$$C = c \iff p(s|C = c, Q = q) > p(s|C = 1 - c, Q = q) \quad (2)$$

Which conditions must Q hold for improving the verification performance with respect to classical verification solutions? The first step in order to address this question is to define the boundary \mathcal{B} between classes $C = 0$ and $C = 1$ in the sq -plane: $\mathcal{B} = \{(s, q) | p(s|C = c, Q = q) = p(s|C = 1 - c, Q = q)\}$. This boundary can also be defined by means of an application that relates each quality measure value q with the score values $\{s_1, \dots, s_p\}$ such that $(q, s_i) \in \mathcal{B}$. Taking into account the definition of verification scores, class 1 (true identity claims) should be more likely for higher values of s , and therefore we can assume that the boundary application associates each quality factor value q with a unique verification score value s . Thus an injective boundary function $\Theta(q)$ can be defined such that $\Theta(q) = s \iff (s, q) \in \mathcal{B}$, and the verification decision can be performed according to the next equation:

$$C = 1 \iff s > \Theta(q) \quad (3)$$

Verification performance of non quality-aided and quality-aided systems can be compared in terms of the Bayesian error. The Bayesian error for a non quality-aided system is:

$$E_{\text{classical}} = \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{\theta} p(s, q|C=1)ds + \int_{\theta}^{+\infty} p(s, q|C=0)ds \right\} dq \quad (4)$$

The Bayesian error for the new framework is defined as:

$$E_{\text{quality}} = \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{\Theta(q)} p(s, q|C=1)ds + \int_{\Theta(q)}^{+\infty} p(s, q|C=0)ds \right\} dq \quad (5)$$

If the threshold application $\Theta(q)$ is optimal in terms of Bayesian error then $E_{\text{classical}} \geq E_{\text{quality}}$, and the equality holds only if $\theta = \Theta(q) \forall q$. Therefore there is a demonstrated theoretical gain whenever the optimal quality-dependent threshold $\Theta(q)$ is not a constant in q . In other words, the necessary and sufficient condition to obtain a performance gain when using a quality factor is that the optimal threshold between classes is not a constant when expressed as a function of this quality factor.

4 Q-based Score Normalization

A solution to determine the threshold as a function of Q is to divide the problem in many simple independent problems. If the Q space \mathcal{E}_Q is divided in a number K of disjoint and connected neighbourhoods \mathcal{N}_i such that $\mathcal{E}_Q = \mathcal{N}_1 \cup \dots \cup \mathcal{N}_K$ and $\mathcal{N}_i \cap \mathcal{N}_j = \emptyset \forall i \neq j$, then it is easy to determine acceptable thresholds θ_i for each neighbourhood \mathcal{N}_i .

This formulation can be easily adapted to the case that many quality measures are provided for one verification modality. In the simple case that only one quality measure is provided, neighbourhoods are intervals defined by their limits $\mathcal{N}_i = (l_i, l_{i+1}]$.

A reasonable approach to build these intervals must take into account that the reliability of any threshold estimation is dependent on the number of verification attempts that belong to that interval. Thus, \mathcal{E}_Q is divided in intervals that contain approximately the same number of verification attempts.

Let us denote the train set as $\mathcal{T} = \{(s_1, q_1, c_1), \dots, (s_{N_T}, q_{N_T}, c_{N_T})\}$, and without generalisation loss, let us suppose they are sorted by the value of the quality factor: $q_i \leq q_j \forall i < j$. If we define a lower bound for the quality measure $q_0 = q_1 - \epsilon$, where ϵ is an arbitrarily small positive constant, then intervals limits can then be defined as:

$$l_i = \frac{1}{2} \left(q_{\lfloor \frac{(i-1)N_T}{K} \rfloor} + q_{\lceil \frac{(i-1)N_T}{K} \rceil} \right) \quad \forall i \in \{1, \dots, K+1\} \quad (6)$$

An optimal threshold θ_i can be easily found for each neighbourhood \mathcal{N}_i . If a soft behaviour of the thresholding function $\Theta(q)$ is assumed, then a good approximation of $\Theta(q)$ can be obtained interpolating the thresholds θ_i . 0th-order

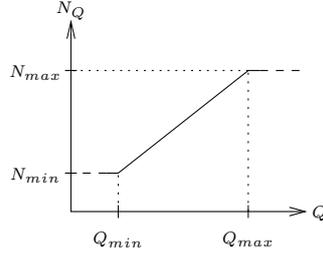


Fig. 1. Number of frames selected as a function of the mean value of the quality measure all along the video sequence.

and 1st-order interpolation close solutions are shown in Sections 4.1 and 4.2 respectively.

The Q-based normalized score is finally computed as $s^Q = s - \Theta(q)$.

4.1 0th-order Threshold Interpolation

0th-order threshold interpolation leads us to a stepwise constant thresholding function that can be formally defined as:

$$\Theta(q) = \begin{cases} \theta_1 & \forall q \leq l_1 \\ \theta_i & \forall q \in \mathcal{N}_i \\ \theta_K & \forall q > l_{K+1} \end{cases} \quad (7)$$

4.2 1st-order Threshold Interpolation

A 1st-order threshold interpolation leads us to a stepwise linear thresholding function that can be formally defined as:

$$\Theta(q) = \begin{cases} 2 \frac{\theta_2 - \theta_1}{l_3 - l_1} q + \frac{\theta_1(l_2 + l_3) - \theta_2(l_1 + l_2)}{l_3 - l_1} & \forall q \leq \frac{l_1 + l_2}{2} \\ 2 \frac{\theta_{i+1} - \theta_i}{l_{i+2} - l_i} q + \frac{\theta_i(l_{i+1} + l_{i+2}) - \theta_{i+1}(l_i + l_{i+1})}{l_{i+2} - l_i} & \forall q \in \left[\frac{l_i + l_{i+1}}{2}, \frac{l_{i+1} + l_{i+2}}{2} \right] \\ 2 \frac{\theta_K - \theta_{K-1}}{l_{K+1} - l_{K-1}} q + \frac{\theta_{K-1}(l_K + l_{K+1}) - \theta_K(l_{K-1} + l_K)}{l_{K+1} - l_{K-1}} & \forall q \geq \frac{l_K + l_{K+1}}{2} \end{cases} \quad (8)$$

5 Quality-based Frame Selection

Quality measures provide information about the reliability of a given score: a verification decision taken on the basis of a high quality frontal face image will be more reliable than a decision taken on the basis of a poor quality frontal face image. This hypothesis will be verified later on Section 8.

A video sequence has not a constant quality. Blurring provoked by fast movements of the user, heavy pose changes, partial occlusions and other factors can affect some frames in the video whilst the other frames can result unaffected.

The idea behind the proposed quality-based frame selection is to keep the most high quality frames in the video for verification, whilst low quality frames are discarded. Besides, a video sequence with a low mean quality measure value will have only a few frames with a high quality, whilst a video with a high mean quality measure value will have many frames with a high quality. Therefore the proposed frame selection keeps the N_Q best frames (those with the highest values of the quality measure) in a video \mathcal{V} , where N_Q grows linearly with the mean quality measure value all along the video $Q_{\mathcal{V}}$. If N_Q is bigger than the number of frames in the video $N_{\mathcal{V}}$, then all the frames in the video are processed. The following equation defines this dependence:

$$N_Q = \begin{cases} N_{min} & \text{if } Q_{\mathcal{V}} < Q_{min} \\ (Q_{\mathcal{V}} - Q_{min}) \frac{N_{max} - N_{min}}{Q_{max} - Q_{min}} + N_{min} & \text{if } Q_{min} \leq Q_{\mathcal{V}} \leq Q_{max} \\ N_{max} & \text{if } Q_{\mathcal{V}} > Q_{max} \end{cases} \quad (9)$$

$$N = \min\{N_Q, N_{\mathcal{V}}\} \quad (10)$$

Figure 1 shows the dependence between N and the mean value of the quality measure Q .

6 Frontal Face Image Quality Measure

The proposed quality measure for frontal face images takes into account two issues:

General image quality The sharpness of the image (associated to a good focus and slow distortion due to fast object movement) is a good general quality measurement. Given an image \mathcal{I} of dimensions $H \times W$, its first order derivative calculated with the Sobel operator $\nabla_{xy}\mathcal{I}$, and its second order derivative calculated with the Laplacian operator $\nabla_{xy}^2\mathcal{I}$, two different coefficients describing the sharpness of the image are derived:

$$\rho_{\text{sob}}(\mathcal{I}) = \frac{\|\nabla_{xy}\mathcal{I}\|}{HW} \quad (11)$$

$$\rho_{\text{lapl}}(\mathcal{I}) = \frac{\|\nabla_{xy}^2\mathcal{I}\|}{HW} \quad (12)$$

Frontal face specific quality The face symmetry is used as a frontal face specific quality factor. Faces with a non frontal pose or with a large rotation will provide a bad symmetry coefficient. Given a frontal face image \mathcal{I} , we define the horizontally flipped version of \mathcal{I} as ${}^f\mathcal{I}$. The asymmetry coefficient of \mathcal{I} is defined as:

$$\rho_{\text{asym}}(\mathcal{I}) = \frac{\|\mathcal{I} - {}^f\mathcal{I}\|}{\|\mathcal{I}\|} \quad (13)$$

However these measures by themselves are not enough to characterise the quality mismatch between the enrolment video and the test video. Let us call

$\widehat{\rho_x^{\text{enrol}}}$ to the mean quality coefficient calculated along the whole enrolment video. Relative quality coefficients are then defined as:

$$\rho_x^{\text{relative}}(\mathcal{I}) = \rho_x(\mathcal{I}) - \widehat{\rho_x^{\text{enrol}}} \quad (14)$$

All the coefficients involved in the frontal face image quality measure are normalized dividing its value by their standard deviation. Finally, the frontal face quality image measure and the frontal face video quality measure are defined as:

$$q_{\mathcal{I}} = - \left[\frac{\rho_{\text{asym}}(\mathcal{I})}{\sigma_{\rho_{\text{asym}}}} + \frac{\rho_{\text{asym}}^{\text{relative}}(\mathcal{I})}{\sigma_{\rho_{\text{asym}}^{\text{relative}}}} \right] + \sum_{x \in \{\text{sob}, \text{lapl}\}} \left[\frac{\log(\rho_x(\mathcal{I}))}{\sigma_{\log(\rho_x)}} + \frac{\rho_x^{\text{relative}}(\mathcal{I})}{\sigma_{\rho_x^{\text{relative}}}} \right]$$

$$q_{\mathcal{V}} = \sum_{\mathcal{I} \in \mathcal{V}} q_{\mathcal{I}} \quad (15)$$

7 Experimental Framework: BANCA Database

The BANCA Database [4] is divided in two disjoint groups $g1$ and $g2$ with 13 males and 13 females each. Each user records 12 video sessions, where one true and one false identity claim is performed. False identity claims are performed always to users with the same gender and in the same group. Sessions are divided in 3 different environments: controlled (good quality recordings), degraded (recordings artificially degraded) and adverse (bad quality recordings). Four sessions are recorded in each environment. The experiments conducted on this paper follow the Pooled protocol defined in [4]. This protocol uses one true identity claim from the controlled conditions for enrolment purposes, and the rest of the database for testing. This protocol provides us with a good quality enrolment and a wide quality range in the test attempts. The Weighted Error Rate (WER) is used for performance measurement:

$$WER(\rho) = \frac{\rho FAR + FRR}{1 + \rho}, \quad (16)$$

where FAR stands for False Acceptance Rate and FRR stands for False Rejection Rate. For each of the values of $\rho = \{1, 10\}$, thresholds are obtained for each group. Then this thresholds are used in the other group and the test WER is obtained. This performance measure allows us to evaluate the system performance in conditions where FAR and FRR must be balanced ($\rho = 1$) or FAR is more critical ($\rho = 10$).

8 Experiments

In the experiments conducted on this paper we test the verification performance of three systems. The first one is the video-based face verification system described in Section 2. This system is used as a baseline. The second system is an

System	Group	$WER(10)$	$WER(1.0)$
Reference System	1	9.09 _(6.34,11.84)	16.72 _(12.54,20.90)
	2	5.23 _(4.33,6.13)	11.28 _(7.45,15.11)
	All	7.14 _(5.69,8.59)	13.98 _(11.09,16.86)
0 th -order Q-norm	1	6.63 _(4.63,8.62)	13.83 _(9.84,17.83)
	2	6.49 _(4.40,8.58)	9.48 _(5.90,13.06)
	All	6.56 _(5.11,8.01)	11.64 _(8.90,14.38)
1 st -order Q-norm	1	6.05 _(4.05,8.05)	14.28 _(10.34,18.22)
	2	7.13 _(4.50,9.76)	8.41 _(5.01,11.81)
	All	6.60 _(4.94,8.27)	11.32 _(8.64,14.00)
0 th -order Q-norm and Q-selection	1	6.25 _(4.44,8.06)	11.83 _(7.91,15.75)
	2	7.56 _(5.18,9.95)	9.27 _(5.73,12.82)
	All	6.92 _(5.42,8.43)	10.54 _(7.89,13.19)
1 st -order Q-norm and Q-selection	1	6.08 _(4.08,8.08)	12.83 _(8.80,16.87)
	2	6.95 _(4.56,9.33)	8.20 _(4.85,11.56)
	All	6.52 _(4.96,8.08)	10.49 _(7.85,13.12)

Table 1. $WER(0.1)$, $WER(1.0)$ and $WER(10)$ face verification performance for the reference system, the Q-based score normalization and the joint Q-based score normalization and frame selection.

improved version of the baseline system: the Q-based score normalization techniques described in Section 4 are incorporated into the baseline system. Finally, the third system is an improved version of the second system: the Q-based frame selection strategy described in Section 5 is incorporated into the second system.

The Q-based normalization system needs one only parameter: the number of quality bins that group the identity claims. After some experiments this number was finally fixed to 4 for convenience. On the other hand, the Q-based frame selection technique requires some parameters to be fixed. These parameters depend on the mean length of the videos to be tested and the range of the quality measure. In our experiments we used $Q_{max} = 0$, $Q_{min} = -20$, $N_{max} = 550$ and $N_{min} = 100$.

Figure 2 shows the true and false score distribution as a function of the frontal face image quality factor described in Section 6. Optimal constant threshold and the a posteriori thresholds obtained for this group by the quality-based 0th-order and 1st-order score normalization techniques are also plotted for comparison purposes.

Table 1 shows the $WER(1.0)$ and $WER(10)$ face verification performance for the three tested systems.

Finally, DET curves for both group 1 and group 2 of BANCA are plotted in Figure 3.

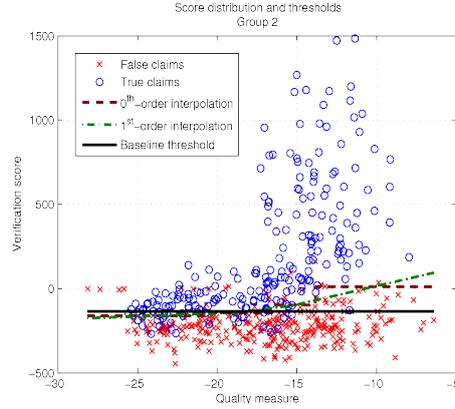


Fig. 2. Score distribution and a posteriori thresholds found for the group 2 of BANCA users.

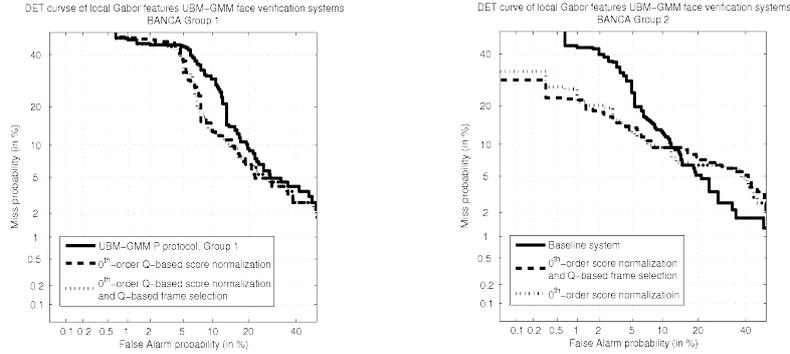


Fig. 3. DET curves of the three video-based face recognition systems for group 1 (left) and group 2 (right) of the BANCA Database.

9 Discussion

Figure 2 shows the thresholds estimated using both 0th-order and 1st-order score normalization techniques and the threshold obtained when the quality measure is not taken into account (the constant black line). It is clear that quality-based approaches lead to non constant class boundaries and therefore the incorporation of this quality measure into the verification process will provide improvements in verification performance. Furthermore, this Figure also shows that high-quality attempts are associated, as hypothesized in Section 5, with larger reliability values than low-quality attempts, where clouds of points belonging to true and false identity claims are more overlapped. This motivates the use of the quality-based frame selection algorithm proposed in this paper.

Experimental results shown in Table 1 show that the incorporation of the frontal face quality measure presented in Section 6 using the Q-based score normalization technique for both 0th-order and 1st-order obtain verification im-

provements in $WER(1.0)$. It also obtains small improvements in $WER(10)$ (a quality measure designed to evaluate biometric systems in high security environments). Most important, statistically significant improvements are obtained in $WER(1.0)$ when the Q-based frame selection algorithm is also incorporated into the verification system. Figure 3 shows that DET curves of quality-aided systems (using the Q-based frame selection or not) are very similar. However, results in Table 1 show that frame selection provides improvements in verification performance. This indicates that a priori thresholds fit better after the frame selection, enabling better WER results.

10 Conclusions

This paper presented a theoretical framework and two practical solutions to incorporate quality measures into any verification process. A quality-based frame selection technique has been also presented. Besides, a new quality measure for frontal face images was presented. This quality aids were incorporated into a GMM-UBM video-based face verification system based on Gabor wavelets. Experiments on the P protocol of the BANCA Database demonstrate the convenience of the proposed face image quality measure and the effectiveness of both quality-based score normalization and quality-based frame selection techniques.

References

1. P. Jonathon Phillips, Patrick Grother, Ross J. Micheals, Duan M. Blackburn, Elan Tabassi, and Mike Bone. Face Recognition Vendor Test. Evaluation Report. Technical report, NISTIR, 2003.
2. José Luis Alba Castro, Daniel González Jiménez, Enrique Argones Rúa, Elisardo González Agulla, and Enrique Otero Muras. Pose-corrected Face Processing on Video Sequences for Webcam-based Remote Biometric Authentication. *Journal of Electronic Imaging*, 17, January 2008.
3. Julián Fierrez Aguilar, Javier Ortega García, Joaquín González Rodríguez, and Josef Bigün. Discriminative Multimodal Biometric Authentication based on Quality Measures. *Pattern Recognition*, 38(5):777–779, 2005.
4. Enrique Bailly-Baillière, Samy Bengio, Frédéric Bimbot, Miroslav Hamouz, Josef Kittler, Johnny Mariéthoz, Jiri Matas, Kieron Messer, Vlad Popovici, Fabienne Porée, Belen Ruiz, and Jean-Philippe Thiran. The BANCA Database and Evaluation Protocol. In *Lecture Notes in Computer Science*, volume 2688, pages 625 – 638, January 2003.
5. Rainer Lienhart and Jochen Maydt. An Extended Set of Haar-like Features for Rapid Object Detection. In *Proceedings of the 2002 International Conference on Image Processing*, volume I, pages 900 – 903, 2002.
6. Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, and Christoph von der Malsburg. Face Recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775 – 779, July 1997.
7. Krzysztof Kryszczuk and Andrzej Drygajlo. Q – stack: Uni- and Multimodal Classifier Stacking with Quality Measures. In Michal Haindl, Josef Kittler, and Fabio Roli, editors, *7th International Workshop on Multiple Classifier Systems*, volume 4472 of *LNCS*, pages 367 – 376, 2007.

Face Quality Assessment System in Video Sequences

Kamal Nasrollahi, Thomas B. Moeslund

Laboratory of Computer Vision and Media Technology, Aalborg University
Niels Jernes Vej 14, 9220 Aalborg Øst, Denmark
{kn, tbm}@cvmt.dk

Abstract. When a person passes by a surveillance camera a sequence of image is obtained. Before performing any analysis on the face of a person, the face first needs to be detected and secondary the quality of the different face images needs to be evaluated. In this paper we present a system based on four simple features including out-of-plan rotation, sharpness, brightness and resolution, to assess the face quality in a video sequence. These features are combined using both a local scoring system and weights. The system is evaluated on two databases and the results show a general agreement between the system output and quality assessment by a human.

Keywords: Face quality assessment, face detection, out-of-plan rotation, surveillance video.

1 Introduction

Considering a person passing by a surveillance camera, a sequence of images of that person is captured by the camera. Depending on the application, most of these images are useless due to problems like not facing the camera, motion blur, darkness and too small size of the region of interest in that image. Usually considering some (one or two) of the best images is sufficient. There is therefore a need for a mechanism which chooses the best image(s) in terms of quality in a sequence of images. This is called Quality Assessment. Image quality assessment is useful in surveillance cameras and also in other applications such as compression, digital photography to inform the user that a low- or high-quality photo had been taken, printing to encourage (or discourage) the printing of better (or poorer) pictures and image management to sort out good from poor photos [1]. This paper is concerned with Face Quality Assessment (FQA).

In different works related to FQA [1-5], different features of the face have been used including: Sharpness, illumination, head rotation, face size, presence of skin pixels, openness of eyes and red eyes. Xiufeng et al. [4] have tried to standardize the quality of face images by facial symmetry based methods. Adam and Robert [5] have extracted 6 features for each face and after assigning a score to each feature, combines them into a general score. Subasic et al. [2] consider more features and interpret the scores related to each feature as a fuzzy value. Fronthaler et al. [3] have studied

orientation tensor with a set of symmetry descriptors to assess the quality of face images.

In a face quality assessment system there are two problems to be dealt with: Reduction the computation and increasing the reliability of the system. In this paper we deal with the first problem by using few and simple features. We have analyzed different features and found that 4 features are sufficient for FQA. These features are out-of-plan-rotation, sharpness, brightness, and face size. In order to deal with the second problem, which is increasing the reliability of the system, we have used locally scoring technique.

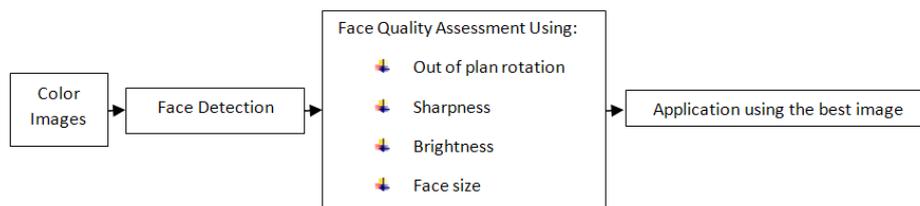


Fig. 1. Block Diagram of the proposed system

The block diagram of the proposed system is shown in figure 1. Given a sequence of color images, the face detection step extracts the face region(s) for each image and feeds them to the face quality assessment block. In this step the quality of the faces for each image is computed and at the end of the sequence the best face for each individual in this sequence is chosen and fed to an application for further processing.

Since some of the features used in the face detection step are also used by the face quality assessment block we briefly describe it in the next section. In Section 3 the assessment process is presented and section 4 shows the experimental results and finally section 5 concludes the paper.

2 Face Detection

Face detection is not the main focus of this paper but since some of the extracted features for the face(s) in this block are used in the assessment process too, we briefly describe it here.

Given a color image, first of all, according to a Gaussian model of skin color, a probability image of the input image is produced. Then using an adaptive threshold, this probability image is segmented to a binary one which has the skin regions separated from the non-skin ones. Here after a cascading classifier using the extracted features for each region decides if this region is a face or not (See figure 2.).

The extracted features for each face include: face size, center of the mass and orientation of the face, number of holes inside the face, the holes area to its surrounding area. For more details regarding the face detection the reader is referred to [6].

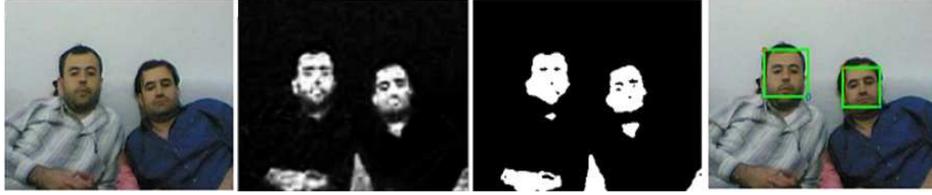


Fig. 2. Face detection process, From left to right: Input color image, its probability and segmented counterpart and detected faces

3 Quality assessment

For each face region detected by the Face Detection, we use both some of the extracted features from the face detection block, and also new features to assess the quality of them. For each feature we assign a locally computed score so that we can decide which image is the best in terms of quality in the given sequence of images. The following subsections describe the details of these features and the scoring process.

3.1 Pose estimation: least out-of-plan rotated face(s)

This feature is one of the most important features in assessing the usability of the face, because wide variation in pose can hide most of the useful features of the face. The previous face quality assessment systems [2, 4, 5] have involved facial features like vertical position of the eyes, distance between the two eyes and vertical symmetry axis to estimate the pose of the face. It is obvious that most of these features may be hidden in various conditions like having spectacles or different lightening condition or even in rotations more than 60° [5]. Hence using the facial features to estimate the pose of the face cannot be reliable. Furthermore in the quality assessment the exact rotation of the face is not important but choosing the least rotated face is. So, we deal with the face as a whole, and calculate the difference between the center of mass and the center of the detected face region. Whenever the rotation of the face increases the difference between these two points increases too.

Given a face in a binary image as shown in figure 3, we calculate the center of mass using the following equation:

$$x_m = \frac{\sum_{i=1}^n \sum_{j=1}^m ib(i, j)}{A}, \quad y_m = \frac{\sum_{i=1}^n \sum_{j=1}^m jb(i, j)}{A} \quad (1)$$

where (x_m, y_m) is the center of mass, b is the binary image containing the detected region as a face, m is the width, n is the height of the detected region and A is the area of this region.

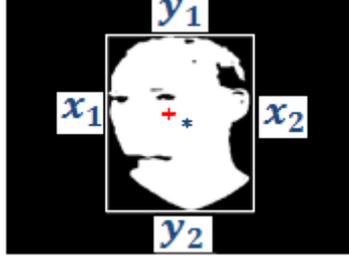


Fig. 3. Center of mass (+) and center of the region (*)

Then we calculate the center of the region detected as a face using the equation 2:

$$x_c = \frac{x_2 - x_1}{2}, y_c = \frac{y_2 - y_1}{2} \quad (2)$$

where x_1 and x_2 are the right most and the left most pixel in the face region and y_1 and y_2 are the lowest and top pixel, respectively, in this region as shown in figure 3. Now we calculate the distance between these two centers as:

$$D = \sqrt{(x_c - x_m)^2 + (y_c - y_m)^2} \quad (3)$$

The minimum value of this distance in a sequence of images gives us the least out-of-plan rotated face as shown in figure 4. To convert this value to a local score in that sequence we use the following equation for each of the images in the sequence:

$$S_1 = \frac{D_{min}}{D} \quad (4)$$

where D_{min} is the minimum value of the D in the given sequence.

Distance	62.26	57.26	46.28	24.82	18.97	15.04
S_1	0.24	0.26	0.32	0.6	0.79	1

Fig. 4. A sequence of different head poses and the associated values for the distance and S_1

Since the center of mass and the detected region are known from the face detection block the only computation for obtaining this feature is equation 3. The technique used by [5, 7] in order to compute this feature, involves the analysis of gradients to locate the left and right sides of face as well as the vertical position of the eyes. From these values the approximate location of the eyes is estimated and the brightest point between the eyes is expected to lie on the face's axis of symmetry. Their method is

not effective when subjects are wearing glasses, or when faces are not close to frontal. While our method is robust in these cases (see the following figure).

					
Distance	89.19	35.55	19.60	24.04	79.22
S_1	0.21	0.55	1	0.81	0.24

Fig. 5. The introduced feature in presence of spectacles and the associated scores

3.2 Sharpness

Since in real world applications the objects are moving in front of the camera, it is possible that the captured image is affected by motion blur, so defining a sharpness feature can be useful for FQA.

					
Sharpness	1.76	1.07	0.99	0.96	0.92
S_2	1	0.6	0.56	0.54	0.52

Fig. 6. An image with different sharpness conditions and the associated scores

Well-focused images, which have a better sharpness compared to blurring images, should get a higher score for this feature. Following [8], if $a(x, y)$ be a part of the image which contains the face, $la(x, y)$ be the result of applying a low-pass filter to it, then the average value of the pixels of the following equation is the sharpness of the face:

$$E = \text{abs}(a(x, y) - la(x, y)) \quad (5)$$

Since it is difficult, at least computationally, to consider an upper limit for the best value of sharpness for all face images in order to have an acceptable normalization, we have used a local maximum. In this way, after calculating the sharpness for all of the chosen faces we assign the following score to the sharpness of each of them:

$$S_2 = \frac{E}{E_{max}} \quad (6)$$

where E_{max} is the maximum value of the sharpness in this sequence. Figure 6 shows some images of one person with different values in sharpness and their associated scores.

3.3 Brightness

Dark images of a face are in general not usable, so we need a way to measure the brightness of the face. Since the region of the face is usually a small region then we can consider the average value of the illumination component of all of the pixels in this region as the brightness of that region. So the following score determines the brightness of the image:

$$S_3 = \frac{B}{B_{max}} \quad (7)$$

where B_{max} is the maximum value of the brightness in this sequence. Figure 7 shows some images of one person in different brightness conditions and their associated scores.

					
Brightness	148.77	138.62	125.74	115.50	111.37
S_3	1	0.93	0.84	0.77	0.74

Fig. 7. An image with different brightness conditions and their associated score

The brighter the image, the higher the score, yields the risk of favoring too bright images. In the real surveillance sequences too bright images are uncommon and in the case of a too bright image a face detector is highly to disregard the face anyway.

3.4 Image resolution

Faces with higher resolution can yield better results than lower resolution ones. But it is only true up to a specific limit [5]. This limit depends on the application which is going to use the face after assessing its quality. But usually considering 50 and 60, respectively, for the width and height of the face is suitable [5, 6]. So we can define the score related to the image resolution as follows:

$$s_4 = \min \left\{ 1, \frac{width}{50} \times \frac{height}{60} \right\} \quad (8)$$

3.5 Choosing the best face in a given sequence

After calculating the four above mentioned features for each of the images in a given sequence, we combine the scores of these features into a general score for each image, as shown in the following equation:

$$S = \frac{\sum_{i=1}^4 w_i s_i}{\sum_{i=1}^4 w_i} \quad (9)$$

where s_i are the score values for the above features and w_i the associated weights for each score. The images are sorted based on their combined scores and depending on the application, one or more images with the greatest values in S are considered as the highest quality image(s) in the given sequence.

4 Experimental Evaluations

We have used both still images and movie samples to evaluate our system. The still images are from the FRI CVL [9] database (DB1). This database consists of sequences of 114 individuals. Each sequence has 7 images with different head rotation (figure 5). Since the images in this database do not have wide variations in sharpness, brightness and size we have used them mainly for assessing the first feature. If the other features of the face have not had wide variations, the least rotated face can give us the best face in terms of the visibility of the facial features.

In order to assess the other features as well as the first feature, we have used the video dataset prepared for the Hermes project [10] (DB2). This dataset contains 48 sequences (6 videos for each of the 8 participants) where these people walk towards a camera while looking from side to side. This provides good examples for assessing all the features together.

According to equation 9 and the experimentally obtained values for the weights of the scores which are shown in Table 1, a combined score is produced for each image in each sequence. The images in each sequence are sorted based on this quality score.

Table 1. The values of the scores weight

<i>Weight</i>	w_1	w_2	w_3	w_4
Value	1	0.9	0.6	0.8

In order to compare the above explained quality scores of the proposed system to a human perception of quality, we have annotated the images in each sequence in our datasets according to their visual features and the visibility of the face and sorted them manually based on our perception of the quality. Table 2 illustrates the results of this comparison using these two databases, in which, the correct matching means the matching between the human perception and system results for the best images in each sequence. While the quality of the images is not too poor and the faces size is not too small the order of the selected images by the proposed system is similar to the order of the selected images by the human. By the way, even for the poor quality images, although it is possible that the images be sorted in different way by the system and the human, but in 100% of the cases we can find the best chosen image by the human inside the first four chosen images by the system.

Table 2. Experimental results

Database	Number of sequences	Number of faces in sequences	Face detection rate	Correct matching
DB1	114	7	94.3%	92.1%
DB2	48	avg. 15	90.5%	87.1%

Figure 8 shows the results of the quality based ranking by the proposed system and the human for an examples from the FRI CVL dataset. In general the human and system rankings are in agreement. Slight differences like those seen in the figure occur when the images in the database are very similar e.g., like the three in the center.

Human ranking	5	4	1	2	3	4	5
System ranking	5	3	1	1	2	4	6

Fig. 8. An example from the FRI CVL database and the quality based rankings

Figure 9 shows a sequence of images from the Hermes dataset and the results of sorting their faces based on the quality both by a human and the proposed system. It is obvious from these images that the selected faces by the system match to the selected faces by the human for the first five images.

Extracted Face								
Human ranking	8	7	6	5	4	3	1	2
System ranking	8	6	7	5	4	3	1	2

Fig. 9. Quality based rankings for a sequence from Hermes dataset

Figure 10 shows another example from the Hermes dataset. In this sequence the size of the head is not changing widely. But since the person turning around his head while walking the other features are changing. It can be seen that in this case the most important feature is head rotation and the proposed system ranking has an acceptable agreement with the human ranking.

							
Extracted Face							
Human ranking	5	3	2	1	6	4	7
System ranking	4	2	3	1	6	5	7

Fig. 10. Quality based rankings in the presence of head rotation

Figure 11 shows another example from the Hermes dataset in which the quality of the images are very poor and the walking person has spectacles. In this figure the details of our locally assigned scores and also the combined scores are shown.

							
Extracted Faces							
Human ranking	7	6	5	4	3	2	1
S_1	0.25	0.14	0.17	0.27	0.71	0.37	1
S_2	0.73	0.75	0.75	0.86	1	0.89	0.85
S_3	0.95	0.94	0.93	0.95	0.97	0.98	1
S_4	0.2	0.28	0.37	0.51	0.64	0.96	1
S	0.49	0.48	0.51	0.61	0.81	0.76	0.95
System ranking	6	7	5	4	2	3	1

Fig. 11. A poor quality sequence of images and the details of the locally scoring technique

As seen in the above figures (8-11), the quality based rankings by the proposed system and the human are very close. A few incorrect ordering were observed due to: our system cannot detect the exact direction of the face, as well as the facial expressions. When the images in the sequence are very similar and the face image are too small then the possibility of miss ranking by the system increases. But in general very good results are obtained.

5 Conclusion

In this paper we present a face quality assessment system based on four simple features including out-of-plan rotation, sharpness, brightness and resolution. These features are combined using both a local scoring system and weights. The system is evaluated on two databases and the results show a general agreement between the system output and quality assessment by a human. For all the sequences of these databases (100%) the best chosen image by the human is one of the first four chosen images by the system and in 89.6% of the cases the first chosen image is the same.

References

1. Huitno Luo, A Training-Based No-Reference Image Quality Assessment Algorithm, International Conference on image Processing (ICIP), pp. 2973--2976, IEEE Press, (2004)
2. M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, Face Image Validation System, In: 4th International Symposium on Image and Signal Processing and Analysis, pp. 30--36, (2005)
3. H. Fronthaler, K. Kollreider, J. Bigun, Automatic Image Quality Assessment with Application in Biometrics, In: International Conference on Computer Vision and Pattern Recognition, IEEE Press, (2006)
4. G. Xiufeng, Z. Stan, L. Rong, P. Zhang, Standardization of Face Image Sample Quality, In: ICB, pp. 242--251, Springer-Verlag, Berlin (2007)
5. A. Fournay, R. Laganiere, Constructing Face Image Logs that are Both Complete and Concise, In: 4th Canadian Conference on Computer Vision and Robot Vision, IEEE Press, Canada (2007)
6. K. Nasrollahi, M. Rahmati, T. B. Moeslund, A Neural Network Based Cascaded Classifier for Face Detection in Color Images with Complex Background, Submitted to ICIAR 2008, Portugal, (2008)
7. K. Peng, L. Chen, S. Ruan, G. Kukharev, A Robust Algorithm for Eye Detection on Gray Intensity Face Without Spectacles, In: Journals of Computer Science and Technology, 5(3), pp. 127--132, (2005)
8. F. Weber, "Some quality measures for face images and their relationship to recognition performance", In: Biometric Quality Workshop. National Institute of Standards and Technology, Maryland, USA, March 8-9 (2006)
9. F. Solina, P. Peer, B. Batagelj, S. Juvan, J. Kovac, "Color-based face detection in the "15 seconds of fame" art installation", In: Conference on Computer Vision / Computer Graphics Collaboration for Model-based Imaging, Rendering, image Analysis and Graphical special Effects, pp. 38--47, France (2003)
10. Hermes project (FP6 IST-027110): <http://www.cvmt.dk/projects/Hermes/index.html>

On quality of quality measures for classification

Krzysztof Kryszczuk and Andrzej Drygajlo

Swiss Federal Institute of Technology Lausanne (EPFL)
Institute of Electrical Engineering, Speech Processing and and Biometrics Group
{krzysztof.kryszczuk, andrzej.drygajlo}@epfl.ch

Abstract. In this paper we provide a theoretical discussion of the impact of uncertainty in quality measurement on the expected benefits resulting from an inclusion of quality measures in classification. While an ideal signal quality measure should be a precise quantification of the actual signal properties relevant to the classification process, real quality measurement may be uncertain. We show how does the degree of uncertainty in quality measurement impact the gains in class separation achieved thanks to using quality measures as conditionally relevant classification feature, and demonstrate that while noisy quality measures become irrelevant, they do not impair class separation beyond the baseline result. We present supporting experimental results using synthetic data.

Key words: quality measures, feature relevance, classifier ensembles

1 Introduction

Degradation of biometric signal quality has been shown to impair the performance of biometric classification systems. One of the remedies to this problem is the use of dedicated metrics that capture the quality of biometric signals. These metrics are referred to as quality measures (*qm*). Ideally, a *qm* should aptly quantify the direct impact that the extraneous, noisy factors have on the collected signal with respect to the deployed classifier. An example of such quality measure could be a microphone that records the noise that masks the speech, but does not capture the speech at the same time. However, it is not always possible or practical to devise a setup capable of capturing the impact of extraneous, quality-degrading factors directly. In this case, one must infer the quality degradation from the collected signals themselves. Indeed, this is frequently the case for most biometric quality measures proposed in the literature. Consequently, an indirect measurement of the quality degradation may carry a measurement error and the *qm* may be to some degree uncertain. An important question arises: how much does the uncertainty in quality measurement impact the value of quality measures as auxiliary feature from the viewpoint of biometric classification? This paper answers this question from a theoretical perspective. We adopt the generic framework of classification with quality measures, *Q-stack*, proposed in [1, 2], since it has been shown to be a generalization of existing algorithms of classification with *qm*. Consequently, the results presented in this paper are valid for

any instantiation of classification system with qm accounted for by the model of $Q - stack$. Using this framework as a reference, we prove that uncertainty in quality measurement reduces the conditional relevance of quality measures as features to the stacked classifier. We demonstrate the practical implications of this finding using synthetic datasets, where additive and multiplicative noise models are used. This paper is structured as follows: Section 2 gives a theoretical discussion of the impact of uncertainty in qm measurement on class separation, Section 3 gives experimental support for the theoretical findings, and Section 4 concludes the paper.

2 Signal quality and quality measures

In a typical biometric classification system with quality measures one has two sources of complementary information: the baseline scores x , and the quality measures qm . The baseline scores x are obtained from biometric classifiers operating on feature sets derived from class-selective raw biometric data, and can be viewed as a compressed representation of this data. The quality measures qm convey information about the conditions of data acquisition and the extent of extraneous noise that shapes the raw data, and therefore are class-independent. In order to make use of the quality information, many algorithms have been proposed - for single classifier systems they were often referred to as *adaptive model/threshold selection* [3], while for multiple-classifier systems they are frequently referred to as *quality-dependent fusion* [4, 5]. Recently proposed framework of $Q - stack$ [1] is a generalization of these methods, where baseline classifier scores x and quality measures qm are features to a second-level stacked classifier which models the dependencies between x and qm . From this perspective, qm become conditionally relevant classification features, which together with scores x grant better class separation than that achieved in the domain of x alone.

If a degradation of observed biometric data is important from the classification perspective, this fact will be reflected in a shift of score x . In real situations quality measures must be estimated from measurement of the noisy process that degrades the data. An ideal quality measure would give an error-free estimate of every instance of noise affecting particular observations and score x . In practice, especially in situations where measuring the noisy process directly is impossible or very troublesome, quality measures must be derived from the observed data itself - and indeed this is the most common case [6]. It is then likely that the quality measurement is done with certain degree of error, or uncertainty.

A schematic representation of this situation is shown in Figure 1. The parameters of processes A , B , N and D as well as the nature of the function $\Phi(x', n)$ are used exclusively for the purpose of data generation and are never used in order to adjust the parameters of classifiers applied. Let us now denote the class separation $D_{A,B}^1$ obtained in the domain of x alone

$$D_{A,B}^x = \int_{-\infty}^{\infty} |p_A(x) - p_B(x)| dx. \quad (1)$$

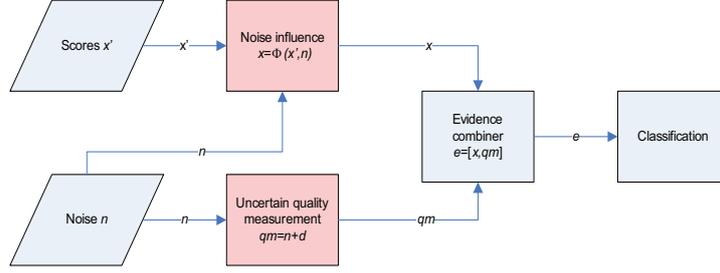


Fig. 1. Diagram of the data flow of the experiment.

Consider class separation $D_{A,B}^e$ in the domain of evidence space $e = [x, qm]$, defined between class conditional distributions $p_A(x, qm)$ and $p_B(x, qm)$. This separation can be expressed in terms of Matusita distance

$$\begin{aligned} D_{A,B}^e &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |p_A(x, qm) - p_B(x, qm)| dx dqm = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |p_A(qm|x)p_A(x) - p_B(qm|x)p_B(x)| dx dqm \end{aligned} \quad (2)$$

Suppose that qm is measuring the actual signal-distorting condition n with uncertainty d , reflected in an increasingly noisy measurement of qm . In this situation

$$d \gg n \Rightarrow p_A(qm|x) = p_B(qm|x) = p(qm), \quad (3)$$

where $p(qm)$ is the stochastic process that describes the observed noisy $qm = n + d$. Consequently

$$\begin{aligned} D_{A,B}^e &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} p(qm) |p_A(x) - p_B(x)| dx \right) dqm = \\ &= \int_{-\infty}^{\infty} p(qm) \left(\int_{-\infty}^{\infty} |p_A(x) - p_B(x)| dx \right) dqm = \\ &= \int_{-\infty}^{\infty} p(qm) D_{A,B}^x dqm = D_{A,B}^x \int_{-\infty}^{\infty} p(qm) dqm = D_{A,B}^x. \end{aligned} \quad (4)$$

The result given by Equation 4 is the same if the expression for Matusita distance is replaced by another distance measure, for instance by divergence or by the Kullback-Leibler distance.

The important conclusion is that independently of the marginal distributions of x and qm or their mutual dependence relationships, increasing the uncertainty in quality measurement results in the reduced conditional relevance of qm . As a result, class separation in the space defined by qm and x approaches class separation observed in the domain of x alone. In Section 3 we demonstrate the implications of this finding using synthetic datasets.

3 Experiments

In this section we illustrate the theoretical predictions given in Section 2. Using synthetic datasets we show the impact of increasing uncertainty in quality measurement on classification accuracy. The choice of synthetic over real biometric datasets is dictated by the fact that in reality measurement of noise without uncertainty is impossible. Also, once quality measurement is taken, there is no way of knowing what this uncertainty actually is. The use of synthetic datasets gives us the experimental comfort of a full control over all stochastic, data-generating processes involved in the experiment.

In the experiments reported here, we generate following data: hypothetical (never observed), noise-free baseline classifier scores x' , Environmental condition n which effects scores x' , resulting in noisy baseline scores $x = \Phi(x', n)$, and measurement noise d , which causes uncertainty in measurement of n . Quality measures qm are estimated according to $qm = n + d$. In the absence of noisy measurement, qm measures n without uncertainty. We use two models $\Phi(x', n)$ of impact of environmental conditions on scores, namely an additive ($\Phi(x', n) : x = x' + n$), and an multiplicative noise model ($\Phi(x', n) : x = x'n$).

We use four different classifier types as stacked classifiers: a Linear Discriminant Analysis - based classifier: *LDA*, a Quadratic Discriminant Analysis - based classifier: *QDA*, a Bayes classifier using Gaussian Mixture Model - based distribution representation: *Bayes*, and a Support Vector Machines - based classifier using RBF kernel: *SVM*. Separate training and testing datasets are generated. The classifiers are trained using 1000 training data points and then deployed to classify another 1000 testing data points. The knowledge of the underlying statistical processes is not used to tune the parameters of the deployed stacked classifiers. For each noise model type, the magnitude of uncertainty in quality measurement d was controlled by adjusting the variance σ_d^2 . Classification performance of the stacked classifiers as a function of correlation ρ between the resulting quality measures qm and noisy scores x was recorded. The value of ρ is a measure of dependence between x and qm which can be evaluated in practical applications.

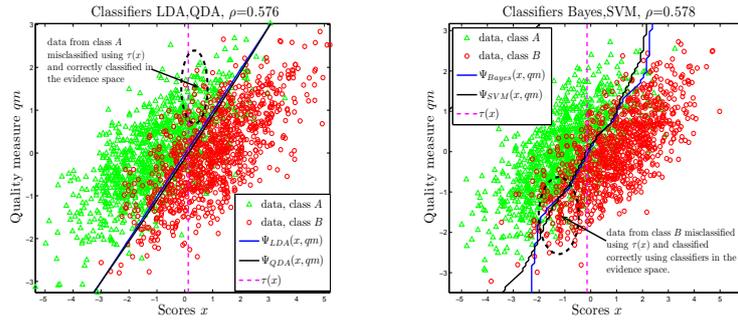
3.1 Additive noise model

Consider Gaussian processes, which generate observations according to $p(x'|A) = \mathcal{N}(\mu_{x',A}, \sigma_{x',A}^2)$ and $p(x'|B) = \mathcal{N}(\mu_{x',B}, \sigma_{x',B}^2)$, where $\mu_{x',A} = -1, \sigma_{x',A}^2 = 1$, and $\mu_{x',B} = 1, \sigma_{x',B}^2 = 1$. Bayes error [7] associated with the classification of x' into classes A and B can be analytically estimated to be $E'_{Bayes} \approx 0.1587$. Let the noise-generating process N produce noise instances n according to $p(n) = \mathcal{N}(\mu_N, \sigma_N)$, $\mu_N = 0, \sigma_N^2 = 1$. If no noise would be present, observed scores would be $x = x'$. Assume that in the presence of noise N the observed scores x are affected by the noise n according to $x = \Phi(x', n) = x' + n$. consequently the class-conditional distributions of observed scores $p(x|A)$ and $p(x|B)$ are given by convolution of the probability density functions [8]: $p(x|A) = p(x' + n|A) =$

$$p(x'|A) * p(n|A) = \mathcal{N}(\mu_N + \mu_{x',A}, \sigma_N^2 + \sigma_{x',A}), \text{ and } p(x|B) = p(x' + n|B) = p(x'|B) * p(n|B) = \mathcal{N}(\mu_N + \mu_{x',B}, \sigma_N^2 + \sigma_{x',M}).$$

Let us now measure the quality measure qm . Of course it would be best to measure n directly, $qm \propto n$. This ideal measurement may in practice be not feasible and the noise measurement may be uncertain. We model this possible uncertainty by adding white Gaussian noise of controlled variance σ_d^2 to the measurement of qm . In this scenario, for $\sigma_d^2 = 0 \Rightarrow qm \propto n$, and for $\sigma_d^2 \rightarrow \infty$ the quality measure qm becomes independent on the actual noise n , and thus it ceases to be informative from the viewpoint of classification using Q -stack. Since all involved processes are Gaussian then the dependency between quality measurements and scores can be measured by computing the correlation coefficient ρ between qm and x .

In the experiments shown in this section we classify 1000 testing data points, using classifiers trained on a separately generated set of 1000 training data points. The data are generated by processes described above. The impact of ρ on the class-conditional distributions evidence, $p(\mathbf{e}|A)$, $p(\mathbf{e}|B)$ is shown in Figure 2.



(a) Classifiers $\tau(x)$, LDA and QDA (b) Classifiers $\tau(x)$, Bayes and SVM

Fig. 2. Class-conditional evidence distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ with Q -stack decision boundaries for LDA, QDA, SVM and Bayes classifiers. Quality measures taken at $\sigma_d^2 = 0$.

In Figure 2 quality measures qm represent the ideal case when, $\sigma_d^2 = 0$. In the experiments shown in Figure 2 this resulted in the correlation coefficient between scores x and quality measures qm of $\rho \approx 0.58$. Corresponding decision boundaries Ψ_{LDA} , Ψ_{QDA} , Ψ_{Bayes} , Ψ_{SVM} are shown, as estimated by corresponding stacked classifiers, as well as the baseline score decision threshold $\tau(x)$.

Figure 3 demonstrates graphically an example of the impact of the uncertainty in estimating qm , on classification results in the evidence space $\mathbf{e} = [x, qm]$. Here, the measurement of quality measure is very noisy at $\sigma_d^2 = 20$, resulting in a low correlation coefficient between x and qm of $\rho \approx 0.13$. Conse-

quently the decision boundaries Ψ between classes A and B tend towards $x = \tau$, the decision boundary obtained when using only x as classification feature. As the difference between classification in the evidence spaces of $\mathbf{e} = [x, qm]$ and $e = [x]$ wanes with growing σ_d^2 , so does the benefit of using quality measure as added dimension in the evidence vector.

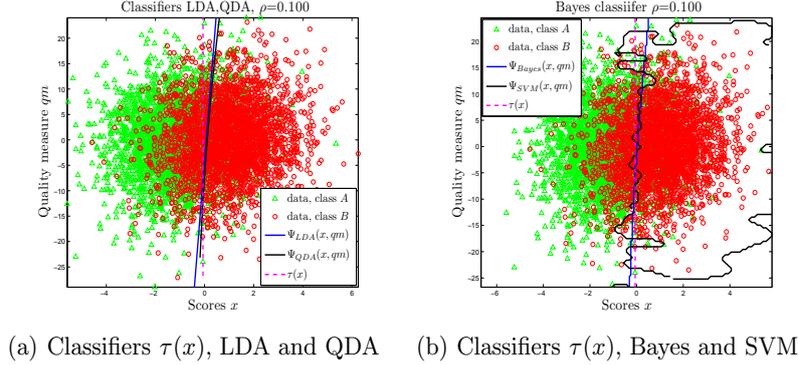


Fig. 3. Class-conditional evidence distributions $p(\mathbf{e}|A)$ and $p(\mathbf{e}|B)$ with Q – stack decision boundaries for LDA, QDA, SVM and Bayes classifiers. Quality measures taken at $\sigma_d^2 = 20$

Compare the behavior of the decision boundary Ψ_{SVM} in Figures 2(b) and 3(b). The curve shown in Figure 3(b) shows a clear overfitting to the training data as a result of an increase in dimensionality of \mathbf{e} beyond necessity. This is not the case in Figure 2(b). Such overfitting may be avoided by clustering quality measures [9], or by simply choosing a classifier of a smaller parametric complexity.

Figure 4 presents the explicit relationship between the correlation coefficient ρ between x and qm and the classification error rates in the evidence space using decision boundaries $\tau(x)$, Ψ_{LDA} , Ψ_{QDA} , Ψ_{Bayes} and Ψ_{SVM} . The variance σ_d^2 of the process D that adds uncertainty to the measurement of qm was changed from $\sigma_d^2 = 0$ to $\sigma_d^2 = 20$. Figure 4 shows the classification errors after 50 independent experimental runs in terms of mean Half Total Error Rate (HTER). The error bars show the standard deviation of HTER. The numerical results of this experiment are gathered in Table 3.1.

3.2 Multiplicative noise model

In this section we show an analogous experiment as reported in Section 3.1, but here the noise N is multiplicative rather than additive. Now, the parameters of stochastic processes A , B and N are $\mu_{x',A} = 3$, $\sigma_{x',A}^2 = 1$, $\mu_{x',B} = 6$, $\sigma_{x',B}^2 = 3$, and $\mu_N = 4$, $\sigma_N^2 = 1$. Noise instances n are affecting x' according to the function

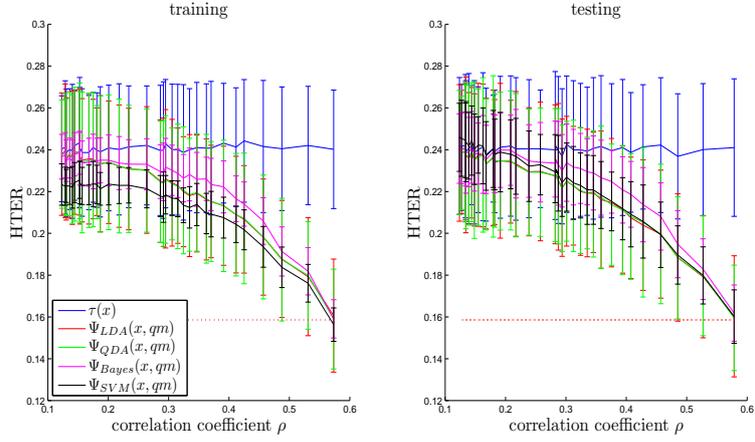


Fig. 4. Impact of the correlation ρ between the observed scores x and the observed quality measures qm , for additive noise.

Table 1. Selected *HTER* results from Figure 4(b), 1000 data points, mean values and standard deviations after 50 repetitions for each value of σ_d^2 .

σ_d^2	0	0.4	1	2.6	7	15	20
$\rho(x, qm)$	0.5785	0.4852	0.4075	0.3012	0.2086	0.1441	0.1272
<i>HTER</i>							
$\mu_{\tau(x)}$	0.241	0.2368	0.2387	0.2426	0.2418	0.2423	0.2411
$\sigma_{\tau(x)}$	0.0329	0.0299	0.0286	0.0326	0.0336	0.0286	0.0302
μ_{LDA}	0.1603	0.1884	0.2074	0.2249	0.2343	0.2386	0.2389
σ_{LDA}	0.029	0.0305	0.0291	0.031	0.0294	0.0312	0.0295
μ_{QDA}	0.1596	0.1883	0.208	0.2249	0.2349	0.2388	0.2385
σ_{QDA}	0.0252	0.0293	0.0301	0.0287	0.0333	0.0333	0.0271
μ_{Bayes}	0.1618	0.1948	0.2186	0.2338	0.2398	0.2411	0.2406
σ_{Bayes}	0.0134	0.016	0.0157	0.0185	0.0168	0.0171	0.0169
μ_{SVM}	0.1602	0.1896	0.2091	0.2269	0.2378	0.2444	0.2455
σ_{SVM}	0.0128	0.0139	0.0156	0.0163	0.0165	0.0165	0.0182

$x = \Phi(x', n) = n \cdot x'$, generating noisy observations (scores) x . Similarly as in Section 3.1, the uncertainty in measuring qm is controlled by adjusting σ_d^2 . Examples of evidence distributions and corresponding decision boundaries are shown in Figure 5 (for $\sigma_d^2 = 0$), and in Figure 6 (for $\sigma_d^2 = 20$)

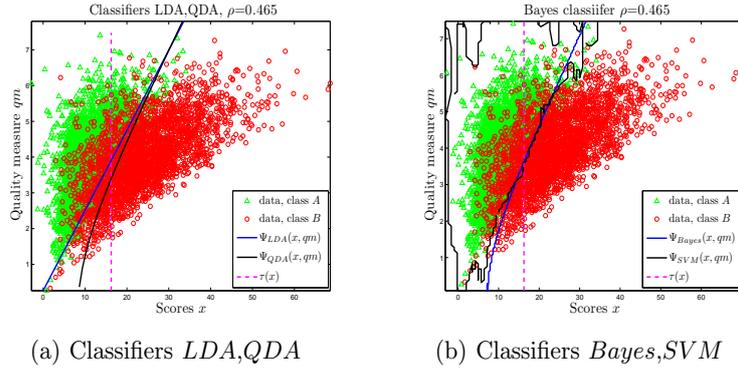


Fig. 5. Classification in the evidence space $\mathbf{e} = [x, qm]$ using (a) *LDA, QDA*, and (b) *Bayes, SVM* stacked classifiers, for $\sigma_d^2 = 0$.

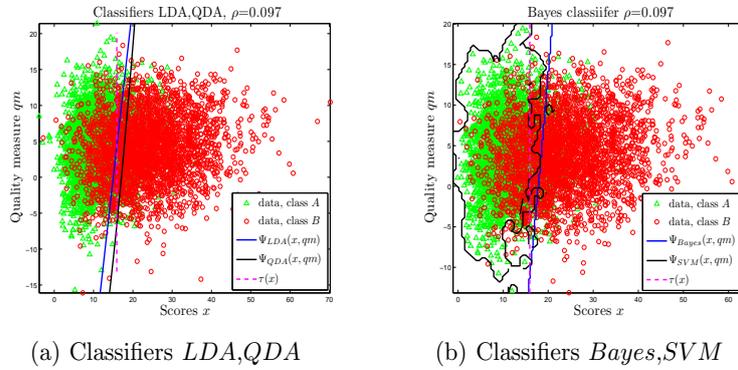


Fig. 6. Classification in the evidence space $\mathbf{e} = [x, qm]$ using (a) *LDA, QDA*, and (b) *Bayes, SVM* stacked classifiers, for $\sigma_d^2 = 20$.

Figure 7 presents the explicit relationship of the correlation coefficient ρ between x and qm and the classification error rates in the evidence space using decision boundaries $\tau(x)$, Ψ_{LDA} , Ψ_{QDA} , Ψ_{Bayes} and Ψ_{SVM} , for σ_d^2 changed in the range of 0 to 20. Obtained classification errors are recorded for respective classifiers after 50 independent experimental runs in terms of mean Half Total

Error Rate (HTER). The error bars show the standard deviation of HTER. Numerical data from this experiment are gathered in Table 2.

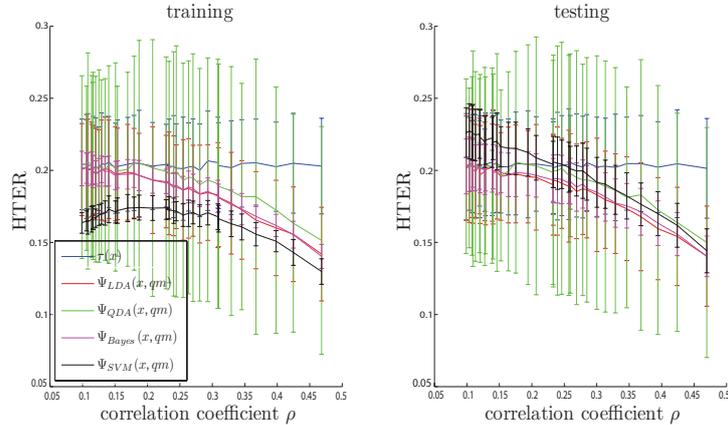


Fig. 7. Impact of the correlation ρ between the observed scores x and the observed quality measures qm , for multiplicative noise.

4 CONCLUSIONS

We have proved that the use of uncertain qm in the framework of Q – *stack* does not impair class separation in the evidence space, and therefore does not negatively impact class separation in respect to the baseline systems which do not use qm . We have instantiated this theoretical result with synthetic datasets, using additive and multiplicative models of noise. The conducted experiments showed that as the uncertainty of qm increases, the classification performance approaches that of a system that uses no quality measures. This result is explained by the fact that uncertain qm lose their conditional relevance to the classification process.

Another important conclusion from the presented study concerns the problem of model selection for classification with qm . As the presented results show, qm collected with a high certainty allow for successful deployment of stacked classifiers of less constrained complexity. As the uncertainty of qm grows, stacked classifiers of restricted complexity proved to be less prone to overtraining than those with more degrees of freedom. This overtraining result can be clearly seen from the error bars in Figures 4 and 7, but also from the shapes of decision boundaries in Figures 2, 3, 5 and 6. This result shows that the sensitivity of the stacked classifier to overtraining depends not only on problem dimensionality but on the strength of dependencies between variables in the evidence vector.

Table 2. Selected *HTER* results from Figure 7(b), 1000 data points, mean values and standard deviations after 50 repetitions for each value of σ_d^2 .

σ_d^2	0.000	0.400	1.000	2.600	7.000	15.000	20.000
$\rho(x, qm)$	0.000	0.007	0.012	0.014	0.019	0.022	0.022
<i>HTER</i>							
$\mu_{\tau(x)}$	0.201	0.202	0.202	0.203	0.202	0.205	0.203
$\sigma_{\tau(x)}$	0.035	0.032	0.037	0.035	0.033	0.030	0.037
μ_{LDA}	0.141	0.160	0.177	0.191	0.197	0.203	0.202
σ_{LDA}	0.035	0.035	0.035	0.031	0.033	0.033	0.036
μ_{QDA}	0.150	0.173	0.186	0.201	0.199	0.203	0.203
σ_{QDA}	0.080	0.086	0.085	0.087	0.069	0.060	0.061
μ_{Bayes}	0.140	0.163	0.177	0.194	0.197	0.203	0.203
σ_{Bayes}	0.014	0.015	0.017	0.017	0.015	0.016	0.019
μ_{SVM}	0.144	0.170	0.186	0.204	0.215	0.225	0.226
σ_{SVM}	0.015	0.015	0.017	0.017	0.018	0.017	0.017

References

1. Kryszczuk, K., Drygajlo, A.: Improving classification with class-independent quality measures: Q-stack in face verification. In: Proc. of the 2nd International Conference on Biometric ICB'07, Seoul, Korea (2007)
2. Kryszczuk, K., Drygajlo, A.: Q-stack: uni- and multimodal classifier stacking with quality measures. In: Proceedings of the International Workshop on Multiple Classifier Systems, Prague, Czech Republic (May 2007)
3. Wein, L., Baveja, M.: Using fingerprint image quality to improve the identification performance of the U.S. VISIT program. In: Proceedings of the National Academy of Sciences, 2005. (2005)
4. Fierrez-Aguilar, J.: Adapted Fusion Schemes for Multimodal Biometric Authentication. PhD thesis, Universidad Politecnica de Madrid (2006)
5. Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.K.: Quality-based score level fusion in multibiometric systems. In: Proceedings of International Conference on Pattern Recognition. Volume 4., Hong Kong, China (August 2006) 473–476
6. Richiardi, J., Kryszczuk, K., Drygajlo, A.: Quality measures in unimodal and multimodal biometric verification. In: Proceedings of the 15th European Conference on Signal Processing EUSIPCO 2007, Poznan, Poland (September 2007)
7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd edn. Wiley Interscience, New York (2001)
8. Grinstead, C.M., Snell, J.L.: Introduction to Probability. American Mathematical Society (1997)
9. Poh, N., Heusch, G., Kittler, J.: On combination of face authentication experts by a mixture of quality dependent fusion classifiers. In: Proceedings of the 7th International Workshop on Multiple Classifier Systems, Prague, Czech Republic (2007)

Definition of Fingerprint Scanner Image Quality Specifications by Operational Quality

A. Alessandroni¹, R. Cappelli², M. Ferrara², D. Maltoni²

¹CNIPA – Centro Nazionale per Informatica nella Pubblica Amministrazione
via Isonzo 21/B, 00198 Roma, Italy. E-mail: alessandroni@cnipa.it

²C.d.L. Scienze dell'Informazione - Università di Bologna, via Sacchi 3, 47023 Cesena, Italy
DEIS – Viale Risorgimento, 2 – 40126 Bologna, Italy.
E-mail: {cappelli, ferrara, maltoni}@csr.unibo.it

Abstract. This paper analyzes two recently released image quality specifications for single-finger scanners and proposes three new specifications targeted to different types of applications. A comparison of the potential effects on fingerprint recognition accuracy of the various specifications is carried out using an approach based on the definition of “operational quality”. The experimental results show that the three new image quality specifications proposed in this work have an accuracy/cost tradeoff better than the existing ones.

1. Introduction

Fingerprint recognition is one of the most reliable and effective biometric technologies and is being adopted as the main identity verification method in several large scale applications. Some countries already store fingerprint data in electronic identity documents and many others plan to do so in the near future. Examples of recent large-scale government projects based on fingerprint recognition include: the US-VISIT [12] and PIV [10] programs in the United States, the Biometric Passport in Europe [2], the Malaysian government multipurpose card [7] and the Singapore biometric passport [11] in Asia.

In large-scale biometric applications, the choice of the acquisition devices is one of the most critical issues since many, often conflicting, requirements have to be taken into account, such as the need for high-quality images, interoperability requisites and budget.

Typically, in large-scale projects a set of specifications is given for the input devices, in order to guarantee a minimum quality level for some

relevant parameters. In the FBI Image Quality Specifications (IQS) for fingerprint scanners [3] [4], the “quality” is defined as “fidelity” in reproducing the original fingerprint pattern, and it is quantified by parameters traditionally used for vision, acquisition and printing systems: geometric accuracy, gray level dynamic range, Signal-to-Noise Ratio (SNR), Spatial Frequency Response (SFR), etc. This definition of quality is clearly appropriate to IAFIS and other applications where the images may be examined by forensic experts. In fact human experts’ comparison techniques heavily rely on very fine details such as pores, incipient ridges, etc., for which the fidelity to the original signal is fundamental.

On the other hand, the situation is different in totally-automated biometric systems, where: i) the images are stored but used only for automated comparisons, or ii) only fingerprint templates are stored. As discussed in a recent work [1], in these cases it may be more appropriate to define the fingerprint scanner quality as the ability of a fingerprint scanner to acquire images that maximize the accuracy of automated recognition algorithms (*operational quality*). A first advantage of the operational quality is that it allows to estimate the loss of performance of a scanner compliant to a given IQS with respect to an “ideal scanner”. In [1], the impact on the recognition accuracy of each quality parameter has been separately assessed, to understand which are the most critical requirements. This work evaluates the simultaneous effect of all the requirements referring to two recently released IQS for single-finger scanners (PIV and PassDEÜV) and proposes three new sets of IQS (CNIPA-A, CNIPA-B and CNIPA-C) targeted to different applications where single finger scanners are required.

The rest of this paper is organized as follows: section 2 reviews and compares the above five fingerprint scanner IQS and section 3 studies their potential impact on recognition accuracy; finally section 4 draws some conclusions.

2. IQS for single finger scanners

This section presents some IQS for single-finger scanners to be used in different applications.

- PIV: established by the US Federal Bureau of Investigation (FBI) for the US Personal Identification Verification program, whose aim is to improve the identification and authentication for access to U.S. Federal facilities and information systems [4] [9];
- PassDEÜV: established by the German Federal Office for Information Technology Security (BSI) for the capture and quality assurance of fingerprints by the passport authorities and the transmission of passport application data to the passport manufacturers [13]; the PassDEÜV requirements are identical to the FBI AFIS requirements (see [3]) except for the acquisition area, which can be smaller;
- CNIPA-A/B/C: these three new set of specifications are here proposed for the first time; they are currently being evaluated by CNIPA (the Italian National Center for ICT in the Public Administration) for inclusion within the guidelines for the Italian public administrations involved in biometric projects. In particular:
 - CNIPA-A is conceived for: i) enrolment in large-scale applications where device interoperability is crucial (e.g. passports, national identity card); ii) identity verification in large-scale applications where the enrolment has been performed with an IAFIS IQS or CNIPA-A compliant scanners (e.g. passport or visa verification);
 - CNIPA-B is conceived for: i) enrolment and verification in medium-scale projects (e.g. intra-organization projects); ii) identity verification in large-scale applications where the enrolment has been performed with CNIPA-A scanners (e.g. national identity card verification);
 - CNIPA-C is conceived for enrolment and verification in small-scale applications, where typically users are authenticated on the same device (e.g. logical and physical security in small organizations).

The five IQS are mainly based on the following quality parameters:

- *Acquisition area*: capture area of the scanner (w×h).
- *Native resolution*: the scanner's true internal resolution (R_N) in pixels per inch (ppi).
- *Output resolution*: the resolution of the scanner's final output fingerprint image in ppi.

- *Gray-level quantization*: number of gray-levels in the final output fingerprint image.
- *Geometric accuracy*: geometric fidelity of the scanner, measured as the absolute value of the difference D , between the actual distance X between two points on a target and the distance Y between those same two points as measured on the output scanned image of that target; PIV and PassDEÜV evaluate this parameters in two different modalities: *Across-bar* (D_{AC}) and *Along-bar* (D_{AL}), see [8] [9] for more details, while CNIPA requires to measure the *Relative difference* ($D_{Rel} = \frac{D}{X}$).
- *Input/output linearity*: the degree of linearity is measured as the maximum deviation D_{Lin} of the output gray levels from a linear least squares regression line fitted between input signal and output gray levels scanning an appropriate target (see[8] [9]).
- *Spatial frequency response*: PIV and PassDEÜV evaluate the SFR using the device Modulation Transfer Function (MTF) measured at each nominal test frequency f , using a continuous-tone sine wave target; CNIPA specifications assess this factor by dividing the acquisition area in $0.25'' \times 0.25''$ regions and measuring, for each region, the *Top Sharpening Index* (TSI), see [5] [6] for more details.
- *Gray level uniformity*: defined as the gray-level differences found in the image obtained by scanning a uniform dark (or light) gray target. This parameter is evaluated by dividing the acquisition area in $0.25'' \times 0.25''$ regions and measuring the differences between: i) the average gray-levels of adjacent rows/columns ($D_{RC}^{dark}, D_{RC}^{light}$), ii) the average gray-level of any region and the gray-level of each of its pixels ($D_{PP}^{dark}, D_{PP}^{light}$); iii) the average gray-levels of any two regions ($D_{SA}^{dark}, D_{SA}^{light}$).
- *Signal-to-noise ratio*: the signal is defined as the difference between the average output gray-levels obtained from acquisition of a uniform light gray and a uniform dark gray target, measuring the average values over independent $0.25'' \times 0.25''$ areas; the noise is defined as the standard deviation of the gray-levels in those areas.
- *Fingerprint gray range*: given a set of scanned fingerprint images, the dynamic range (DR) of each image is defined as the total number of gray levels that are present in the image.

Table 1 reports, for each of the above quality parameters, the requirements that a scanner has to meet in order to be compliant with the five specifications.

Table 1. A comparison of PIV, PassDEÜV and CNIPA-A/B/C requirements for the main quality parameters.

Parameter	Requirement				
	PIV IQS [4] [9]	PassDEÜV IQS [13]	CNIPA		
			IQS A	IQS B	IQS C
Acquisition area	$w \geq 12.8\text{mm}$ $h \geq 16.5\text{mm}$	$w \geq 16.0\text{mm}$ $h \geq 20.0\text{mm}$	$w \geq 25.4\text{mm}$ $h \geq 25.4\text{mm}$	$w \geq 15.0\text{mm}$ $h \geq 20.0\text{mm}$	$w \geq 12.8\text{mm}$ $h \geq 16.5\text{mm}$
Native resolution	$R_N \geq 500\text{ppi}$				
Output resolution	$R_N \pm 2\%$	$R_N \pm 1\%$	$R_N \pm 1\%$	$R_N \pm 1.5\%$	$R_N \pm 2\%$
Gray-level quantization	256 gray-levels (8 bpp)				
Geometric accuracy	In 99% of the tests: $D_{AC} \leq \max\{0.0013'', 0.018 \cdot X\}$ $D_{AL} \leq 0.027''$	In 99% of the tests: $D_{AC} \leq \max\{0.0007'', 0.01 \cdot X\}$ $D_{AL} \leq 0.016''$	In all the tests: $D_{Rel} \leq 1.5\%$	In all the tests: $D_{Rel} \leq 2.0\%$	In all the tests: $D_{Rel} \leq 2.5\%$
Input/output linearity	No requirements	$D_{Lin} \leq 7.65$	No requirements		
Spatial frequency response	$MTF_{\min}(f) \leq MTF(f) \leq 1.12$ see [1] for PIV $MTF_{\min}(f)$	$MTF_{\min}(f) \leq MTF(f) \leq 1.05$ see [1] $MTF_{\min}(f)$ values	For each region: $TSI \geq 0.20$	For each region: $TSI \geq 0.15$	For each region: $TSI \geq 0.12$
Gray level uniformity	In 99% of the cases: $D_{RC}^{dark} \leq 1.5$; $D_{RC}^{light} \leq 3$ For 99% of the pixels: $D_{PP}^{dark} \leq 8$; $D_{PP}^{light} \leq 22$ For every two small areas: $D_{SA}^{dark} \leq 3$; $D_{SA}^{light} \leq 12$	In 99% of the cases: $D_{RC}^{dark} \leq 1$; $D_{RC}^{light} \leq 2$ For 99.9% of the pixels: $D_{PP}^{dark} \leq 8$; $D_{PP}^{light} \leq 22$ For every two small areas: $D_{SA}^{dark} \leq 3$; $D_{SA}^{light} \leq 12$	No requirements		
Signal-to-noise ¹	$SNR \geq 70.6$	$SNR \geq 125$	$SNR \geq 70.6$	$SNR \geq 49.4$	$SNR \geq 30.9$
Fingerprint gray range	For 80% of the images: $DR \geq 150$	$DR \geq 200$ for 80% images; $DR \geq 128$ for 99% images	For 10% of the images: $DR \geq 150$	For 10% of the images: $DR \geq 140$	For 10% of the images: $DR \geq 130$

3. Impact of the IQS on the recognition accuracy

In order to evaluate the impact on fingerprint recognition accuracy of the IQS described in section 2, a systematic experimentation has been carried out. Following the testing methodology introduced in [1] and using the same test database, fingerprint images acquired by hypothetical scanners compliant with each IQS have been simulated.

¹ Actually in PIV IQS and CNIPA this requirement is given by setting the maximum noise standard deviation to 3.5. To make it comparable with the corresponding PassDEÜV IQS, here we provide this value as a SNR under the hypothesis of a 247 gray-level range (see [3]): $SNR = 247/3.5 = 70.6$.

To this purpose, the transformations described in [1] have been sequentially applied to the original fingerprint images according to the worst-case scenario hypothesized in table 2.

Table 2. The table reports, for each quality parameter, the characteristic of the scanners hypothesized for enrolment and verification. In fact, in a typical large-scale application, the scanner used during enrolment may be different from those used during verification. Note that “different” does not necessarily imply a distinct model/vendor: in fact, two scanners of the same model may produce different output images. For instance if a certain scanner model is compliant to a $500\text{ppi}\pm 1\%$ output resolution specification, one of such devices may work at 505ppi and another at 495ppi.

Parameter	Enrolment scanner	Verification scanner
Acquisition area	The minimum-allowed	The minimum-allowed
Output resolution	The minimum-allowed ($Res_{OR}-R_{Res}\%$)	The maximum-allowed ($Res_{OR}+R_{Res}\%$)
Geometric accuracy	Negligible	The maximum-allowed
Spatial frequency response	The minimum-allowed	The minimum-allowed
Signal-to-noise ratio	The minimum-allowed	The minimum-allowed
Fingerprint gray range	The minimum-allowed	The minimum-allowed

The outcome of this analysis is an estimation of the loss of accuracy that scanners compliant with each specification may cause with respect to the performance that would be obtained using “ideal” scanners (i.e. devices with negligible perturbations). The loss of accuracy is quantified by the relative EER difference between the two cases, expressed as a percentage value (see [1]); for instance, if the relative EER difference is 100%, it means that the EER obtained by the simulated scanners is twice the EER obtained by the ideal scanners. All the experiments have been carried out using ten state-of-the-art fingerprint recognition algorithms. Figure 1 reports a box-plot for each specification: each box-plot shows descriptive statistics about the relative EER difference of the ten algorithms.

In order to better understand the results summarized in figure 1, it is useful to compare the five IQS as shown in table 3, where the “strictness” of the various quality parameters with respect to the FBI IAFIS IQS [3] is highlighted. The most “tolerant” specification is CNIPA-C, which has the least demanding requirements for all the

parameters: as it was reasonable to expect, this specification can cause the largest performance drop (182% on the average). Less tolerant but still not very strict are PIV and CNIPA-B (both with three “L” and three “M” requirements); however the loss of performance that can be caused by them is definitely different: on the average 156% and 44%, respectively. This means that the impact of the various quality parameters on the recognition accuracy is not uniform: the first three parameters in table 3 are more critical than the last three ones. The two most demanding specifications (PassDEÜV and CNIPA-A) cause definitely smaller performance drops (on the average 20% and 18%, respectively); table 3 shows that CNIPA-A has the most strict requirement for the acquisition area, while PassDEÜV for spatial frequency response, signal-to-noise ratio and fingerprint gray range. CNIPA-A IQS produces the smallest loss of performance, mainly due to the larger acquisition area that is the most critical parameter, as proved in [1].

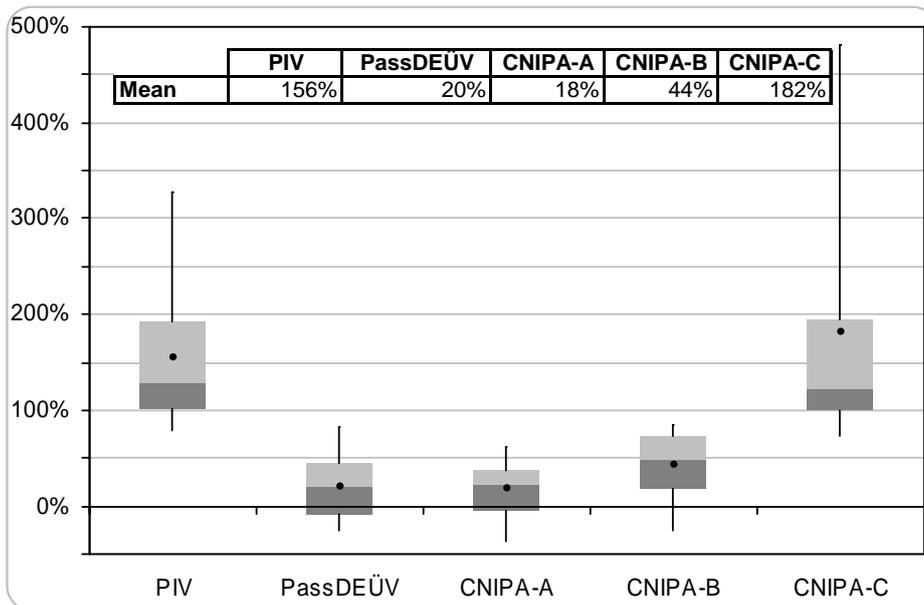


Fig. 1. A box-plot for each specification. Each box-plot graphically shows descriptive statistics of a set of data: the top and bottom of the vertical line denotes the largest and smallest observation, respectively; the rectangle contains 50% of the observations (from the first to the third quartile) and highlights the median (second quartile); finally the mean of all the observations is marked with a black circle.

Table 3. For each of the quality parameters a label in {"L: Low", "M: Medium", "H: High"} is used to characterize the level of "strictness" of the requirement in the specifications. "H" is used when the constraint is as "strict" as in the FBI IAFIS-IQS [3]; "M" and "L" are used when the specification is moderately or significantly relaxed, respectively, with respect to the corresponding FBI IAFIS-IQS.

Parameter	Level of "strictness" of the requirements				
	PIV IQS	PassDEÜV	CNIPA-A	CNIPA-B	CNIPA-C
Acquisition area	L	M	H	M	L
Output resolution accuracy	L	H	H	M	L
Geometric accuracy ²	L	H	H	M	L
Spatial frequency response ³	M	H	M	L	L
Signal-to-noise ratio	M	H	M	L	L
Fingerprint gray range	M	H	M	L	L

4. Conclusions

This paper analyzed two recently released IQS for single-finger scanners (PIV and PassDEÜV) and proposed three new IQS (CNIPA-A/B/C) targeted to different applications. A comparison of the potential effects on recognition accuracy of the various specifications has been carried out using the operational quality approach introduced in [1].

The three new IQS have been designed according to outcomes of [1], and trying to define IQS with an optimal accuracy/cost tradeoff.

Although the results of this analysis partially depend on the specific scanner used for collecting the test database (see [1]), we believe that similar results would be obtained starting from images acquired by other scanners. According to the experimental results, we can conclude that the three proposed specifications are well suited for the applications they are targeted to. In particular:

- CNIPA-A specification is able to guarantee the best performance among the five IQS reviewed, thanks to the higher acquisition area, which proved to be the most important parameter;

² CNIPA-A/B/C IQS set requirements on a slightly different measurement of geometric accuracy; however it can be shown that PIV IQS is comparable to CNIPA-C requirement and PassDEÜV requirement (the same of the IAFIS IQS) is comparable to CNIPA-A requirement (see [1]).

³ Although CNIPA-A/B/C IQS on spatial frequency response are based on a different measure (see [5] [6]), according to our internal tests, PIV-IQS requirement is close to CNIPA-A.

- CNIPA-B specification is able to guarantee an accuracy that is clearly better than PIV and not too far from PassDEÜV; on the other hand, the cost of a device compliant to CNIPA-B would be definitely lower than that of one compliant to PassDEÜV, thanks to the less demanding requirements on five parameters;
- CNIPA-C specification can guarantee an accuracy similar to PIV but, also in this case, the cost of a device compliant to CNIPA-C would be definitely lower than the cost of PIV-compliant devices.

6. Bibliography

- [1] R. Cappelli, M. Ferrara and D. Maltoni, "On the Operational Quality of Fingerprint Scanners", to appear on IEEE Transactions on Information Forensics and Security.
- [2] Council of EU, "Council Regulation (EC) No 2252/2004 of 13 December 2004 on standards for security features and biometrics in passports and travel documents issued by Member States", in: Official Journal of the EU of Dec. 29, 2004, Vol. L 385, pp. 1-6.
- [3] Department of Justice, F.B.I., "Electronic Fingerprint Transmission Specification", CJIS-RS-0010 (V7), January 1999.
- [4] FBI, CJIS Division, "Image Quality Specifications for Single Finger Capture Devices", version 071006, 10 July 2006; download at: <http://www.fbi.gov/hq/cjisd/iafis/piv/pivspec.pdf>, January 2007.
- [5] M. Ferrara, A. Franco and D. Maltoni, "Estimating Image Focusing in Fingerprint Scanners", in proceedings Workshop on Automatic Identification Advances Technologies (AutoID07), Alghero, Italy, pp.30-34, June 2007.
- [6] M. Ferrara, A. Franco and D. Maltoni, "Fingerprint scanner focusing estimation by Top Sharpening Index", in proceedings 14th International Conference on Image Analysis and Processing (ICIAP07), Modena, Italy, pp.223-228, September 2007.
- [7] GMPC Project web site, <http://www.jpn.gov.my/kppk1/Index2.htm>, February 2008.
- [8] N.B. Nill, "Test Procedures for Verifying IAFIS Image Quality Requirements for Fingerprint Scanners and Printers", MITRE Technical Report MTR 05B0000016, April 2005.

- [9] N.B. Nill, “Test Procedures for Verifying Image Quality Requirements for Personal Identity Verification (PIV) Single Finger Capture Devices”, MITRE Technical Report MTR 060170, December 2006.
- [10] PIV Program web site, <http://csrc.nist.gov/piv-program>, February 2008.
- [11] Singapore Biometric Passport web site, <http://app.ica.gov.sg>, February 2008.
- [12] US-VISIT Program web site, <http://www.dhs.gov/us-visit>, February 2008.
- [13] BSI, “Technical Guideline for production data acquisition, quality testing and transmission for passports - Annex 2 – (Version 2.1) Quality requirements for the acquisition and transmission of fingerprint image data as biometric feature for electronic identification documents”, available online at <http://www.bsi.de/english/publications/techguidelines/tr03104>, February 2008.

Modeling Marginal Distributions of Gabor Coefficients: Application to Biometric Template Reduction

Daniel González-Jiménez and José Luis Alba-Castro *

Signal Theory and Communications Department
University of Vigo, Spain
e-mail: {danisub,jalba}@gts.tsc.uvigo.es

Abstract. Gabor filters have demonstrated their effectiveness in automatic face recognition. However, one drawback of Gabor-based face representations is the huge amount of data that must be stored. One way to reduce space is to quantize Gabor coefficients using an accurate statistical model which should reflect the behavior of the data. Statistical image analysis has revealed one interesting property: the non-Gaussianity of marginal statistics when observed in a transformed domain (like Discrete Cosine Transform, wavelet decomposition, etc.). Two models that have been used to characterize this non-normal behavior are the Generalized Gaussian (GG) and the Bessel K Form densities. This paper provides an empirical comparison of both statistical models in the specific scenario of modeling Gabor coefficients extracted from face images. Moreover, an application for biometric template reduction is presented: based on the underlying statistics, compression is first achieved via Lloyd-Max algorithm. Afterwards, only the best nodes of the grid are preserved using a simple feature selection strategy. Templates are reduced to less than 2 Kbytes with drastical improvements in performance, as demonstrated on the XM2VTS database.

1 Introduction

Gabor filters are biologically motivated convolution kernels that have been widely used in face recognition during the last decade (see [1] for a recent survey). Basically, Gabor-based approaches fall into one of the following categories: **a**) Extraction of Gabor responses from a set of *key* points in face images and **b**) Convolution of the whole image with a set of Gabor filters. As highlighted in [1], one of the main drawbacks of these approaches (specially the ones included in category **b**) is the huge amount of memory that is needed to store a Gabor-based representation of the image. Even in the case of **a**), considering 100 points, 40 Gabor filters and float (4 bytes) representation, the template size reaches 32

* This work has been partially supported by Spanish Ministry of Education and Science (project PRESA TEC2005-07212), by the Xunta de Galicia (project PGIDIT05TIC32202PR)

Kbytes which is considerably bigger than those employed by commercial systems. For instance, Cognitec’s [2] templates occupy 1800 bytes each one, and L-1 Identity Solutions’ [3] template size ranges from 648 bytes to 7 Kbytes. One way to reduce the room needed for storing a Gabor-based face representation is to quantize Gabor coefficients using an accurate statistical model.

Statistical analysis of images has revealed, among other characteristics, one interesting property: the non-Gaussianity of image statistics when observed in a transformed domain, e.g. wavelet decomposition. This means that the coefficients obtained through such transformations are quite non-Gaussian being characterized by high kurtosis, sharp central cusps and heavy tails. Among others, the works in [4–7] have observed this behavior, taking advantage of such a property for different applications. Different statistical priors have been proposed to model marginal distributions of coefficients, such as Generalized Gaussians (GGs, pioneered by the work of [4]), Bessel K forms (BKFs) [8] and alpha-stable distributions [9]. In [10], the authors concluded that Bessel K forms were more accurate than the classical Generalized Gaussian densities for modeling marginal distributions. The first goal of this paper is to provide an empirical evaluation of these two priors in the specific context of (Gabor-based) face recognition. Once demonstrated that GGs perform better in this scenario, we took advantage of the underlying statistics to compress data using coefficient quantization by means of Lloyd-Max algorithm. At this point, and in order to further reduce the template size, we decided to apply feature selection by means of the Best Individual Feature (BIF) algorithm [11–13]. This way, the template is compressed because of the lower number of features that are kept and, at the same time, system performance is drastically increased.

The paper is organized as follows: Section 2 presents the system used to extract Gabor features from face images. Section 3 introduces the two statistical densities, Generalized Gaussians and Bessel K Forms, involved in the evaluation, as well as the obtained results. The application for biometric template size reduction based on feature selection and coefficient quantization is presented in Section 4. Finally, conclusions are outlined in Section 5.

2 Gabor Feature Extraction

A set of 40 Gabor filters $\{\psi_m\}_{m=1,2,\dots,40}$ with the same configuration as in [14] (5 spatial frequencies and 8 orientations), is used to extract textural information from face images. The baseline face recognition system that we have used in this paper relies upon extraction of Gabor responses at each of the nodes from a $n_x \times n_y$ (10×13) rectangular grid (Figure 1). All faces were geometrically normalized -so that eyes and mouth are in fixed positions-, cropped to a standard size of 150x116 pixels and photometrically corrected by means of histogram equalization and local mean removal. The region surrounding each grid-node in the image is encoded by the convolution of the image patch with these filters, and the set of responses is called a jet, \mathcal{J} . Therefore, a jet is a vector with 40 *complex* coefficients, and it provides information about a specific region of the

image. At node $\mathbf{p}_i = [x_i, y_i]^T$ and for each Gabor filter ψ_m , $m = 1, 2, \dots, 40$, we get the following Gabor coefficient:

$$g_m(\mathbf{p}_i) = \sum \sum I(x, y) \psi_m(x_i - x, y_i - y) \quad (1)$$

where $I(x, y)$ represents the photometrically normalized image patch. Hence, the complete jet extracted at \mathbf{p}_i is given by $\mathcal{J}(\mathbf{p}_i) = [g_1(\mathbf{p}_i), g_2(\mathbf{p}_i), \dots, g_{40}(\mathbf{p}_i)]$. For a given a face with $n = n_x \times n_y$ grid-nodes $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n\}$, we get n Gabor jets $\{\mathcal{J}(\mathbf{p}_1), \mathcal{J}(\mathbf{p}_2), \dots, \mathcal{J}(\mathbf{p}_n)\}$.

3 Modeling Marginal Distributions of Gabor coefficients

Generalized Gaussians have been already used in [7] to model Gabor coefficients extracted from face images, with good results. On the other hand, [10] compared BKF against GG, concluding that the BKF density fits the data at least as well as the Generalized Gaussian, and outperforms GGs in capturing the heavy tails of the data histogram. The goal of this section is to introduce Generalized Gaussians and Bessel K Forms, and compare the fitting provided by both models in the specific scenario we are considering.

3.1 Univariate Generalized Gaussians

Pioneered by the work of [4], Generalized Gaussians have been successfully used to model marginal distributions of coefficients produced by various types of transforms [5, 6, 15, 7]. The *pdf* of a GG is given by the following expression:

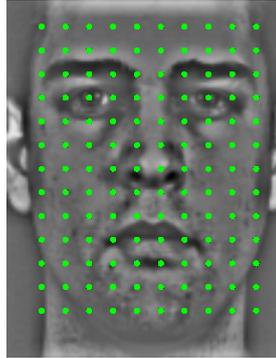


Fig. 1. Rectangular grid over the preprocessed (geometrically and photometrically normalized) face image. At each node, a Gabor jet with 40 coefficients is computed and stored.

$$P_{\mu,\beta,\sigma} = \frac{1}{Z(\beta)\sigma A(\beta)} \exp\left(-\left|\frac{x-\mu}{\sigma A(\beta)}\right|^\beta\right) \quad (2)$$

where β is the so-called *shape* parameter, μ represents the mean of the distribution, and σ is the scale parameter. In the following we will consider zero mean data, i.e. $\mu = 0$. $Z(\beta)$ and $A(\beta)$ in Eq. (2) are given by:

$$Z(\beta) = \frac{2}{\beta} \Gamma\left(\frac{1}{\beta}\right) \quad (3)$$

$$A(\beta) = \sqrt{\frac{\Gamma(1/\beta)}{\Gamma(3/\beta)}} \quad (4)$$

where $\Gamma(\cdot)$ represents the Gamma function. It should be noted that the Laplacian, Gaussian and Uniform distributions are just special cases of this generalized *pdf*, given by $\beta = 1$, $\beta = 2$ and $\beta \rightarrow \infty$ respectively.

3.2 Bessel K Form Densities

Bessel K Form (BKF) densities [8] have recently emerged as a valid alternative for coefficient modeling. As well as the GG, the BKF distribution is characterized by two parameters (p and c) with analogous meaning to that of β and σ respectively. The BKF density is given by:

$$BKF(x; p, c) = \frac{2}{Z(p, c)} |x|^{(p-0.5)} K_{(p-0.5)}\left(\sqrt{\frac{2}{c}} |x|\right) \quad (5)$$

where K_ν is the modified Bessel function of order ν defined in [16], and Z is the normalizing constant given by:

$$Z(p, c) = \sqrt{\pi} \Gamma(p) (2c)^{(0.5p+0.25)} \quad (6)$$

The BKF density is based on a physical model for image formation (the so-called transported generator model), and its parameters have been usually estimated using moments [8], and k statistics unbiased cumulants estimators [10].

3.3 Comparing GGs and BKF's for Modeling Gabor Coefficients of Face Images

As stated above, [10] claims that BKF's outperform GGs. However, no description of the method used to estimate the Generalized Gaussian parameters was included (moments, Maximum Likelihood, etc.). This Section introduces the experimental framework used for evaluating both GGs and BKF's:

From a set of face images $\{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_T\}$, we extract Gabor jets using the rectangular grid introduced in Section 2. Regardless of the node from which they

have been computed, the coefficients corresponding to a given Gabor filter ψ_m (real and imaginary parts separately) are stored together forming two sets of coefficients \mathcal{S}_m^{real} and \mathcal{S}_m^{imag} (Only experiments with the real part are provided. Analogous results were observed for the imaginary part). Now, our goal is to assess which statistical model, GGs or BKF, provides a more accurate fit. To this aim, we performed the following experiment:

- For each pair of orientation and scale, i.e. for each filter ψ_m , both BKF and GG parameters were estimated on 10 different random subsets sampled from \mathcal{S}_m^{real} .
- For each coefficient and set, the Kullback-Leibler (KL) distance [17] was measured between the observed histogram and the two estimated densities.
- The average KL for the m -th coefficient, as well as the associated standard deviation, were stored.

The k statistics unbiased cumulants estimators [10] were used to determine the parameters of the BKF distributions, while Maximum Likelihood (ML) [6] was employed to estimate GG parameters. Examples of observed histograms on a log scale along with the two fitted densities are shown in Figure 2 for coefficients 1, 9, 17, 25 and 33 (i.e. the coefficients with vertical orientation from each frequency subband). From these plots, it seems that both densities are equivalent in the last 3 (lowest) frequency subbands. However, Generalized Gaussians are quite more accurate than BKF in the first two (highest) frequency subbands (specially when fitting the central cusp). In agreement with [10], Bessel K Forms seem slightly better in capturing the heavy tails of the observed histogram for the 1st frequency subband.

Figure 3 shows, for each Gabor coefficient, the mean KL distance (left) as well as the associated standard deviation (right) between the observed histograms and the two estimated densities. It is clear that Generalized Gaussians provide a much better modeling than BKF in the two first scales (highest frequency scales-coefficients from 1 to 16), a slightly better behavior in the third scale (coefficients from 17 to 24) and equal performance in the remaining two scales.

As stated above, BKF parameters were estimated using a robust extension of the moments method, while GG parameters were determined using ML. In [6] it is also described a way to estimate Generalized Gaussian parameters using moments. In order to compare BKF and GGs with similar parameter estimation procedures, the experiment described above was repeated using GGs fitted via the moments-based method. Results are shown in Figure 4, demonstrating that even with comparable estimation procedures, GGs do outperform BKF.

4 Biometric Template Reduction

We have demonstrated that, in the case of Gabor coefficients extracted from face images, the Generalized Gaussians model provides a better fit than the one based on Bessel K Forms. Using the GG model, coefficients can be compressed by means of Lloyd-Max quantization algorithm (the one with minimum mean squared error

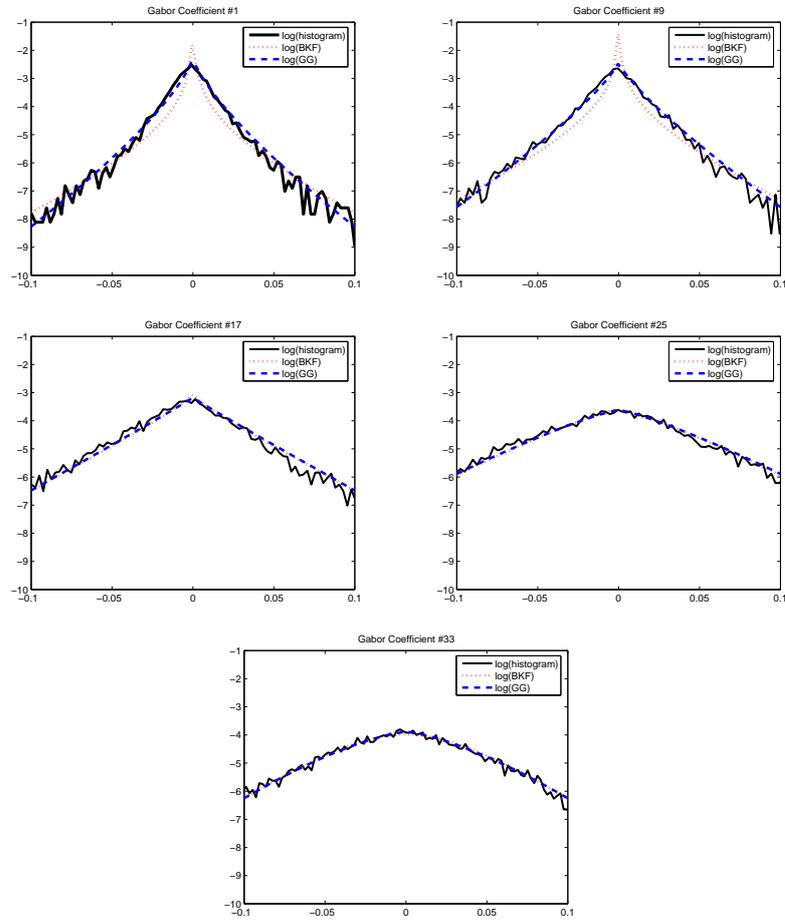


Fig. 2. Examples of observed histograms (on a log scale) along with the BKF and GG fitted densities

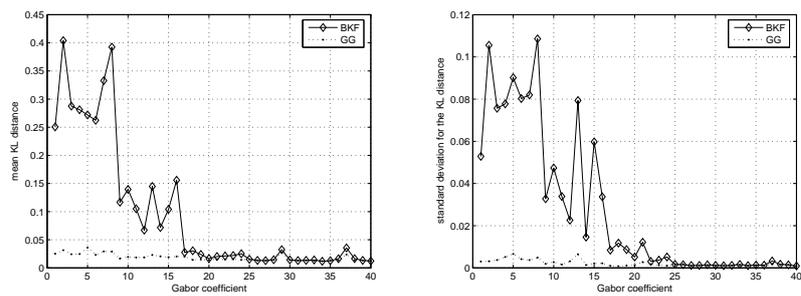


Fig. 3. Left: Mean KL distance between observed histograms and the two estimated densities (GG and BKF). Right: Associated standard deviation

(MSE) for a given number N_L of representative levels) [18, 19]. Hence, instead of storing the original coefficient, we only need to keep two indices (one for the real part and another for the imaginary part) per coefficient ($2 \times 40 \times n$ indices per face). Using N_L quantization levels, we can represent a coefficient with $2 \times \lceil \log_2(N_L) \rceil$ bits. In our case, a face is therefore represented by

$$\frac{40 \times n \times 2 \times \lceil \log_2(N_L) \rceil}{8} = 10n \times \lceil \log_2(N_L) \rceil = 1300 \times \lceil \log_2(N_L) \rceil \text{ bytes} \quad (7)$$

If N_f faces are to be stored, then

$$1300 \times N_f \times \lceil \log_2(N_L) \rceil + 40 \times 4(\text{bytes}) \times N_L(\text{centroids}) \text{ bytes} \quad (8)$$

are needed. The second term in the previous expression represents the storage required for the N_L centroids in each coefficient band (given that both real and imaginary parts have very similar GG parameters, only N_L centroids have been used to quantize each band). Authentication experiments on configuration I [20] of the XM2VTS database [21] demonstrate that high compression rates can be achieved without loss of performance (see Table 1). This table presents the False Acceptance Rate (FAR), the False Rejection Rate (FRR) and the Total Error Rate (TER=FAR+FRR) using both original and compressed (with N_L quantization levels) data. It seems clear that only $N_L = 8$ levels are enough, since no degradation is observed. However, even in this case, 6 bits per coefficient are needed (3 bits for the real part and 3 bits for the imaginary part), and $\frac{130 \times 6 \times 40}{8} = 3900$ bytes are required to store a template. At this point, one can think of, at least, two other possibilities to reduce the amount of data to be stored: *i*) reduce the number of Gabor jets extracted from the rectangular grid and *ii*) reduce the number of coefficients that form a jet. Regarding the former, one straightforward way to achieve the goal is to reduce the size of the grid, but this can lead to a decrease in performance. A wiser strategy may include feature selection, i.e. preserve those jets that are good at discriminating between clients and impostors, and discard the remaining ones. The benefits of

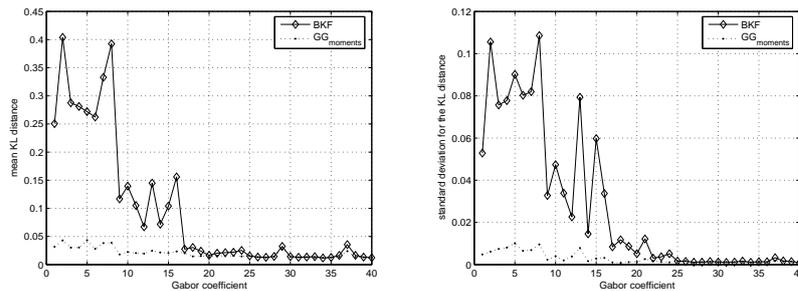


Fig. 4. Left: Mean KL distance between observed histograms and the two estimated densities (GG fitted via a moments-based method and BKF). **Right:** Associated standard deviation

Table 1. Face Verification on the XM2VTS database. False Acceptance Rate (FAR), False Rejection Rate (FRR) and Total Error Rate (TER) over the test set using both raw and compressed data and the whole set of 130 jets. Moreover, approximate storage saving is provided for each quantization level

	Storage Saving	Test Set		
		FAR(%)	FRR(%)	TER(%)
$N_L = 2$	$\approx 97\%$	12.15	18.25	30.40
$N_L = 4$	$\approx 94\%$	4.19	8.00	12.19
$N_L = 8$	$\approx 91\%$	3.49	5.50	8.99
$N_L = 16$	$\approx 87\%$	3.85	5.50	9.35
$N_L = 32$	$\approx 84\%$	3.71	5.00	8.71
$N_L = 64$	$\approx 81\%$	3.53	5.50	9.03
$N_L = 128$	$\approx 78\%$	3.57	5.00	8.57
$N_L = 256$	$\approx 75\%$	3.63	4.75	8.38
$N_L = 512$	$\approx 72\%$	3.66	4.75	8.41
Raw data	0%	3.79	5.25	9.04

such methodology are not limited to reducing storage but also increasing system performance. One technique that has demonstrated good performance despite its simplicity is the Best Individual Feature (BIF) selection approach [11–13]. In [13], different tools for Gabor jet similarity fusion were evaluated, concluding that, in the specific scenario of face verification with little amount of data for building client templates, simple approaches such as BIF performed even better than more complex techniques like SVMs, Neural Networks, etc. (see [13] for details). The idea behind BIF (as its name reads) is to select the best individual features according to some criterion (e.g. individual classification accuracy). We fixed the number of features to be selected by BIF to 1, 10, 20, . . . and performed authentication experiments on configuration I of the Lausanne protocol [20]. The “best” jets were selected employing both training and evaluation data, while system performance was measured on the disjoint test set (see [20] for details on data partition and protocol). Table 2 shows the TER over the test set for different quantization levels and number of jets selected by BIF, highlighting those configurations that achieve less than 5% of TER. By employing the best 20 jets with coefficients quantized using $N_L = 64$ levels, the error rate drops to **4.7%** and, at the same time, template size is reduced to **1.2** Kbytes (see Table 3 for the corresponding template sizes). Equal performance (4.8%) is achieved with original data and 10 jets but at the cost of a considerably bigger (3.2 Kbytes) template size. Compared to the original system with 130 jets and original coefficients, the use of 20 jets and 64 levels implies an increase in performance of 48.09% while approximately saving 97% of space.

Table 2. Face Verification on the XM2VTS database. Total Error Rate (TER) over the test set for different quantization levels and number of jets selected by BIF (N_{Jets})

N_{Jets}	N_L										raw
	2	4	8	16	32	64	128	256	512	data	
1	64.2	42.8	27.2	30.9	23.1	22.3	21.2	21.3	20.8	18.7	
10	37.5	14.1	5.9	5.8	6.5	5.1	5.8	5.4	5.3	4.8	
20	31.4	10.1	5.7	5.6	5.6	4.7	4.8	5.0	4.9	5.3	
30	27.9	10.5	6.6	5.5	5.3	5.4	6.1	5.1	5.3	5.3	
40	25.4	9.1	5.6	5.6	5.7	6.1	5.6	5.4	5.8	5.3	
50	25.4	8.3	6.1	5.8	6.0	5.4	4.9	5.7	5.3	5.4	
60	23.9	8.9	6.5	5.8	6.0	5.6	5.6	5.5	5.9	5.8	

Table 3. Template syze (Kbytes) for different quantization levels and number of jets selected by BIF (N_{Jets})

N_{Jets}	N_L										raw
	2	4	8	16	32	64	128	256	512	data	
1	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.32	
10	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	3.2	
20	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	6.4	
30	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	9.6	
40	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	12.8	
50	0.5	1	1.5	2	2.5	3	3.5	4	4.5	16	
60	0.6	1.2	1.8	2.4	3	3.6	4.2	4.8	5.4	19.2	

5 Conclusions

This paper has presented an empirical comparison of two statistical priors for modeling Gabor coefficients extracted from face images. The main conclusion is that Generalized Gaussians provide a more accurate fitting than Bessel K Forms in this specific scenario. Taking advantage of the underlying statistics, Gabor coefficients were compressed using Lloyd-Max quantization algorithm, and further storage reduction was achieved by means of Best Individual Feature selection. Finally, both biometric template reduction and drastic increase in performance compared to the original system have been obtained.

References

1. Linlin Shen and Li Bai. A Review on Gabor Wavelets for Face Recognition. *Pattern Analysis and Applications*, 9(2):273 – 292, 2006.
2. Cognitec. Cognitec Systems GmbH, <http://www.cognitec-systems.de/>, 2002.
3. L-1. L-1 identity solutions, <http://www.l1id.com/>, 2005.

4. S.G. Mallat. A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.
5. J.R. Hernández, M. Amado, and F. Pérez-González. DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure. *IEEE Transactions on Image Processing*, 9(1):55 – 68, 2000.
6. Minh N. Do and M. Vetterli. Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and Kullback-Leibler Distance. *IEEE Transactions on Image Processing*, 11(2):146 – 158, 2002.
7. D. González-Jiménez, F. Pérez-González, P. Comesaña-Alfaro, L. Pérez-Freire, and J.L. Alba-Castro. Modeling Gabor Coefficients via Generalized Gaussian Distributions for Face Recognition. In *International Conference on Image Processing*, 2007.
8. Anuj Srivastava, Xiuwen Liu, and Ulf Grenander. Universal Analytical Forms for Modeling Image Probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1200–1214, 2002.
9. Alin Achim, Anastasios Bezerianos, and Panagiotis Tsakalides. Novel Bayesian Multiscale Method for Speckle Removal in Medical Ultrasound Images. *IEEE Transactions on Medical Imaging*, 20(8):772–783, 2001.
10. Mohamed-Jalal Fadili and Larbi Boubchir. Analytical form for a Bayesian wavelet estimator of images using the Bessel K form densities. *IEEE Transactions on Image Processing*, 14(2):231–240, 2005.
11. P. Pudil, F. J. Ferri, J. Novovicová, and J. Kittler. Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions. In *Proceedings ICPR*, volume 2, pages 279–283, 1994.
12. Berk Gökberk, M. Okan Irfanoglu, Lale Akarun, and Ethem Alpaydin. Optimal Gabor Kernel Selection for Face Recognition. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2003)*, volume 1, pages 677 – 680, Barcelona, Spain, September 2003.
13. D. González-Jiménez, E. Argones-Rúa, J.L. Alba-Castro, and J. Kittler. Evaluation of Point Selection and Similarity Fusion Methods for Gabor Jets-Based Face Verification. *IET Computer Vision*, 1(3–4):101–112, 2007.
14. L. Wiskott, J. M. Kruger, and C. von der Malsburg. Face recognition by Elastic Bunch Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775 – 779, 1997.
15. K. Sharifi and A. Leon-Garcia. Estimation of Shape Parameter for Generalized Gaussian Distributions in Subband Decompositions of Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1):52–56, 1995.
16. M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1970.
17. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
18. S.P. Lloyd. Least Squares Quantization in PCM. Technical report, Bell Laboratories, 1957.
19. J. Max. Quantizing for Minimum Distortion. *IRE Transactions on Information Theory*, IT-6:7–12, 1960.
20. J. Luttin and G. Maitre. Technical Report RR-21: Evaluation Protocol for the Extended M2VTS Database (XM2VTSDB). Technical report, IDIAP, 1998.
21. K. Messer, J. Matas, J. Kittler, J. Luetin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. *Audio- and Video-Based Biometric Person Authentication*, pages 72 – 77, March 1999.

Bosphorus Database for 3D Face Analysis

Arman Savran¹, Neşe Alyüz², Hamdi Dibeklioglu², Oya Çeliktutan¹, Berk Gökberk³, Bülent Sankur¹, Lale Akarun²

¹Boğaziçi University, Electrical and Electronics Engineering Department
arman.savran, oya.celiktutan, bulent.sankur@boun.edu.tr

²Boğaziçi University, Computer Engineering Department
nese.alyuz, hamdi.dibeklioglu, akarun@boun.edu.tr

³Philips Research, Eindhoven, The Netherlands
berk.gokberk@philips.com

Abstract. A new 3D face database that includes a rich set of expressions, systematic variation of poses and different types of occlusions is presented in this paper. This database is unique from three aspects: i) the facial expressions are composed of judiciously selected subset of Action Units as well as the six basic emotions, and many actors/actresses are incorporated to obtain more realistic expression data; ii) a rich set of head pose variations are available; and iii) different types of face occlusions are included. Hence, this new database can be a very valuable resource for development and evaluation of algorithms on face recognition under adverse conditions and facial expression analysis as well as for facial expression synthesis.

1. Introduction

In recent years face recognizers using 3D facial data have gained popularity due to their lighting and viewpoint independence. This has also been enabled by the wider availability of 3D range scanners. The 3D face processing can be envisioned as a single modality biometric approach in lieu of the 2D version or in a complementary mode in a multi-biometric scheme. Another goal application of 3D facial data is the understanding of facial expressions in an affective human-computer interface.

Most of the existing methods for facial feature detection and person recognition assume frontal and neutral views only, and hence biometry systems have been adapted accordingly. However, this may be uncomfortable for the subjects and limit the application domains. Therefore, the newly emerging goal in this field is to develop algorithms working with natural and uncontrolled behaviour of subjects. A robust identification system can also cope with the subjects who try to eschew being recognized by posing awkwardly and worse still, by resorting to occlusions via dangling hair, eyeglasses, facial hair and other accessories.

On the other hand, understanding of facial expressions has wide implications ranging from psychological analysis to affective man-machine interfaces. Once the expression is recognized, this information can also be used to help the person identifier.

The desiderata of a 3D face database enabling a range of facial analysis tasks ranging from expression understanding to 3D recognition are the following: i) Action units from Facial Action Coding System (FACS) [1], both single and compound; ii) Emotional expressions; iii) Ground-truthed poses; iv) Occlusions originating from hair tassel, eyeglasses and a gesticulating hand. Motivated by these exigencies, we set out to construct a multi-attribute 3D face database.

1.1. Comparisons with Major Open 3D Face Databases

Various databases for 3D face recognition and occasionally 3D expression analysis are available. Most of them are focused on recognition; hence contain a limited range of expressions and head poses. Also, none of them contain face occlusions. One of the most popular 3D database FRGC v.2 [2], though the biggest one in the number of subjects has only a few mild expressions. The database richest in the spectrum of emotional expressions is BU-3DFE [3]. Every subject displays four intensity levels of the six emotions. Table I lists publicly available databases of relevance and compares with our database.

Table 1. List of some well known 3D face databases. Subj.: subjects Samp.: samples per subject, Occl.: occlusions, NA: not available

Database	Subj	Samp.	Total	Expression	Pose	Occl.
FRGC v.2 [2]	466	1-22	4007	Anger, happiness, sadness, surprise, disgust, puffy	NA	NA
BU-3DFE [3]	100	25	2500	Anger, happiness, sadness, surprise, disgust, fear (in 4 levels)	NA	NA
ND2006 [4]	888	1-63	13450	Happiness, sadness, surprise, disgust, other	NA	NA
York [5]	350	15	5250	Happiness, anger, eyes closed, eye-brows raised	Uncontrolled up & down	NA
CASIA [6]	123	15	1845	Smile, laugh, anger, surprise, closed eyes	NA	NA
GavabDB [7]	61	9	549	Smile, frontal accentuated laugh, frontal random gesture	Left, right, up, down	NA
3DRMA [8]	120	6	720	NA	Slight left/right & up/down	NA
Bosphorus	81	31-53	3396	34 expressions (action units & six emotions)	13 yaw, pitch & cross rotations	4 occlusions (hand, hair, eyeglasses)

A new comprehensive multi-expression, multi-pose 3D face database enriched with realistic occlusions is presented in this paper. The database has the following merits: i) in addition to the basic six emotional expressions subjects have acted several action units from the FACS [1]; ii) Various ground-truthed head poses are available; iii) A number of facial occlusion types are captured from the subjects. Finally in order to achieve more natural looking expressions, we have employed actors and actresses from professional theatres, opera and the conservatory school.

The content of the database is given in Section 2, and data acquisition is explained in Section 3. In Section 4 the acquired data are evaluated. Finally conclusion is given in Section 5.

2. Database Content

The database consists of 81 subjects in various poses, expressions and occlusion conditions. Many of the male subjects have facial hair like beard and moustache. The majority of the subjects are aged between 25 and 35. There are 51 men and 30 women in total, and most of the subjects are Caucasian. There are total of 3396 face scans. Each scan has been manually labelled for 24 facial landmark points such as nose tip, inner eye corners, etc, provided that they are visible in the given scan. These feature points are given in Table II.

The database has two versions:

- Bosphorus v.1: This version includes 34 subjects with 10 expressions, 13 poses, four occlusions and four neutral faces, thus resulting in a total of 31 scans per subject.
- Bosphorus v.2: This version is designed for both expression understanding and face recognition. There are 47 people with 53 different face scans per subject. Each scan is intended to cover one pose and/or one expression type, and most of the subjects have only one neutral face, though some of them have two. Totally there are 34 expressions, 13 poses, four occlusions and one or two neutral faces. In addition, Bosphorus v.2 also incorporates 30 professional actors/actresses out of 47, which hopefully provide more realistic or at least more pronounced expressions.

In the following subsections, the collected facial expressions, head poses and occlusions are explained.

2.1. Facial Expressions

Two types of expressions have been considered in the Bosphorus databases. In the first set, the expressions are based on action units (AUs) of the FACS [1]. AUs are assumed to be building blocks of expressions, and thus they can give broad basis for facial expressions. Since each action unit is related with the activation of a distinct set of muscles, they can be assessed quite objectively. Although there are 44 AUs in general, we have collected a subset which consists of those AUs that are easier to

enact. The selected action units were grouped into 20 lower face AUs, five upper face AUs and three AU combinations.

Table 2. Manually labeled 24 facial landmark points.

1. Outer left eye brow	
2. Middle of the left eye brow	
3. Inner left eye brow	
4. Inner right eye brow	
5. Middle of the right eye brow	
6. Outer right eye brow	
7. Outer left eye corner	
8. Inner left eye corner	
9. Inner right eye corner	
10. Outer right eye corner	
11. Nose saddle left	
12. Nose saddle right	
13. Left nose peak	
14. Nose tip	
15. Right nose peak	
16. Left mouth corner	
17. Upper lip outer middle	
18. Right mouth corner	
19. Upper lip inner middle	
20. Lower lip inner middle	
21. Lower lip outer middle	
22. Chin middle	
23. Left ear lobe	
24. Right ear lobe	

In the second set, facial expressions corresponding to certain emotional expressions were collected. These are: happiness, surprise, fear, sadness, anger and disgust. It is stated that these expressions are universal among human races [9].

During acquisition of each action unit, subjects were given explanations about these expressions and they were given feedback if they did not enact correctly. Also to facilitate the instructions, a video clip showing the correct facial motion for the corresponding action unit is displayed on the monitor [10]. However, in the case of emotional expressions, there were no video or photo guidelines so that subjects tried to improvise. Only if they were able to enact, they were told to mimic the expression in a recorded video. Moreover, a mirror was placed in front of the subjects in order to let them check themselves.

In Table III., the 34 expressions in the database are given. Also, Fig. 1 shows some 3D faces displaying the happiness emotions of actors/actresses. These facial images are rendered with texture mapping and synthetic lighting.



Fig. 1. Some samples from happiness expression captured from actors/actresses. Texture mapping and synthetic lighting is applied for rendering.

Table 3. Expressions in the Bosphorus database. Presence of the corresponding expressions are denoted by bullets (•). A sample image for each expression is shown at the bottom part.

Expressions	Scan No	Explanation	v.2	v.1
Lower AUs	1	Lower Lip Depressor - AU16	•	
	2	Lips Part - AU25	•	
	3	Jaw Drop - AU26	•	
	4	Mouth Stretch - AU27	•	•
	5	Lip Corner Puller - AU12	•	•
	6	Left Lip Corner Puller - AU12L	•	
	7	Right Lip Corner Puller - AU12R	•	
	8	Low Intensity Lip Corner Puller - AU12LW	•	
	9	Dimpler - AU14	•	
	10	Lip Stretcher - AU20	•	
	11	Lip Corner Depressor - AU15	•	
	12	Chin Raiser - AU17	•	
	13	Lip Funneler - AU22	•	
	14	Lip Puckerer - AU18	•	
	15	Lip Tightener - AU23	•	
	16	Lip Presser - AU24	•	
	17	Lip Suck - AU28	•	•
	18	Upper Lip Raiser - AU10	•	
	19	Nose Wrinkler - AU9	•	•
	20	Cheek Puff - AU34	•	•
Upper AUs	21	Outer Brow Raiser - AU2	•	•
	22	Brow Lowerer - AU4	•	•
	23	Inner Brow Raiser - AU1	•	
	24	Squint - AU44	•	
	25	Eyes Closed - AU43	•	•
Combined AUs	26	Jaw Drop (26) + Low Intensity Lip Corner Puller	•	
	27	Lip Funneler (22) + Lips Part (25)	•	•
	28	Lip Corner Puller (12) + Lip Corner Depressor (15)	•	
Emotions	29	Happiness	•	•
	30	Surprise	•	
	31	Fear	•	
	32	Sadness	•	
	33	Anger	•	
	34	Disgust	•	

2.2. Head Poses

Various poses of the head are acquired for each subject (Table IV. and Fig. 2). There are three types of head poses which correspond to seven yaw angles, four pitch angles, and two cross rotations which incorporate both yaw and pitch. For the yaw rotations, subjects align themselves by rotating the chair on which they sit to align with stripes placed on the floor corresponding to various angles. For pitch and cross rotations, the subjects are required to look at marks placed on the walls by turning their heads only (i.e., no eye rotation). Thus, a coarse approximation of rotation angles can be obtained.

2.3. Oclusions

For the occlusion of eyes and mouth, subjects choose a natural pose for themselves; for example, as if they were rubbing their eyes or as if they were surprised by putting their hands over their mouth. Second, for the eyeglass occlusion, subjects used different eyeglasses from a pool. Finally, if subjects' hair was long enough, their faces were also scanned with hair partly occluding the face (Table V.).

Table 4. Head poses and occlusions in the Bosphorus database.

Head Poses	
<p>1) Yaw Rotations</p> <p>a) Neutral</p> <p>b) +10°</p> <p>c) +20°</p> <p>d) +30°</p> <p>e) +45°</p> <p>f) +90°</p> <p>g) -45°</p> <p>h) -90°</p> <p>2) Pitch Rotations</p> <p>i) Strong upwards</p> <p>j) Slight upwards</p> <p>k) Slight downwards</p> <p>l) Strong downwards</p> <p>3) Cross Rotations</p> <p>m) Yaw and pitch 1 (approximately 20° pitch and 45° yaw)</p> <p>n) Yaw and pitch 2 (approximately -20° pitch and 45° yaw)</p>	 <p>a b c d e f g</p> <p>h i j k l m n</p>
Oclusions	
<p>a) Occlusion of eye with hand – as natural as possible</p> <p>b) Occlusion of mouth with hand – as natural as possible</p> <p>c) Eye glasses (not sunglasses, normal eyeglasses)</p> <p>d) Hair</p>	 <p>a b c d</p>

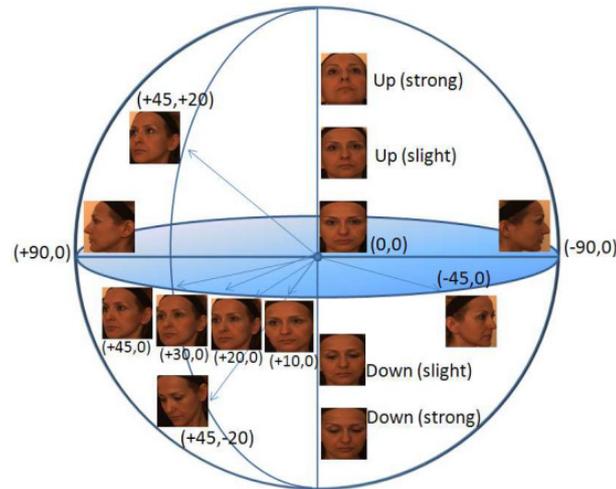


Fig. 2. Head poses in the Bosphorus database.

3. Data Acquisition

Facial data are acquired using Inspeck Mega Capturor II 3D, which is a commercial structured-light based 3D digitizer device [11]. The sensor resolution in x, y & z (depth) dimensions are 0.3mm, 0.3mm and 0.4mm respectively, and colour texture images are high resolution (1600x1200 pixels). It is able to capture a face in less than a second. Subjects are made to sit at a distance of about 1.5 meters away from the 3D digitizer. A 1000W halogen lamp was used in a dark room to obtain homogeneous lighting. However, due to the strong lighting of this lamp and the device's projector, usually specular reflections occur on the face. This does not only affect the texture image of the face but can also cause noise in the 3D data. To prevent it, a special powder which does not change the skin colour is applied to the subject's face. Moreover, during acquisition, each subject wore a band to keep his/her hair above the forehead to prevent hair occlusion, and also to simplify the face segmentation task.

Scanner software is used for acquisition and 3D model reconstruction. The reconstruction from the acquired image data is performed during the acquisition session right after the scanning. This process involves some automatic and manual steps via scanner's software. Although somewhat time consuming, it guarantees that faulty acquisitions are detected and hence can be repeated. In this phase data is also segmented manually by selecting a polygonal face region. In order to remove noise, several basic filtering operations (like Gaussian and Median filtering) are applied. Finally, each scan is down-sampled and saved in two separate files that store colour photograph and 3D coordinates. A segmented 3D face approximately consists of 35K points.

4. Characteristics of the 3D Face Data

In the sequel, we will discuss the pose, expression and occlusion modalities of the face as well as evaluating its quality aspects.

4.1. Discussion of Data Content

This database contains great amount of variations for each individual due to expressions, head poses and occlusions, as explained in Section 2. Important characteristics of these variations are discussed below.

Expressions: Not all subjects could properly produce all AUs, some of them were not able to activate related muscles or they could not control them. Therefore, in the database few expressions are not available for some of the subjects. Also, the captured AUs need to be validated by trained AU experts. Second, since no video acquisition was possible for this database, the AUs were captured at their peak intensity levels, which were judged subjectively. Notice that there was no explicit control for the valence of pronounced expressions. As in any other database, acted expressions are not spontaneous and thoroughly natural. All these factors constitute the limitations of this database for expression studies.

Poses: Although various angles of poses were acquired, they are only approximations. Especially poses including pitch rotations can be subject dependent, since subjects are requested to look at marks placed in the environment. This introduces slight angle differences due to the difference of rotation centres changing from subject to subject. Eye rotation may also cause some difference, though the subjects were warned in that case.

Occlusions: The subject to subject variation of occlusions is more pronounced as compared to expression variations. For instance, while one subject occludes his mouth with the whole hand, another one may occlude it with one finger only; or hair occlusion on the forehead may vary a lot in tassel size and location.

4.2. Discussion of Data Quality

Quality of the acquired data can be quite important depending on the application. Due to 3D digitizing system and setup conditions significant noise may occur. To reduce noise, we tried to optimize experimentally the acquisition setup by trying different lighting conditions and controlling the camera and subject distances. However, there are other sources of problems. These are explained below.

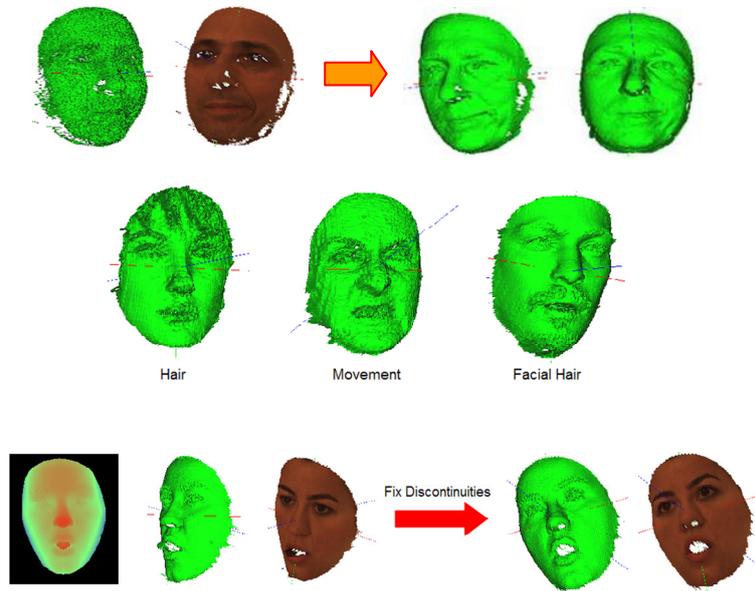


Fig. 3. Commonly occurring problems during image acquisition and face reconstruction. Top row shows basic filtering and self-occlusion problem. In the middle row, noise due to hair, movement, and facial hair is seen. At the bottom left, a mistake in the depth level of the tongue, and at the right, its correction is displayed.

Movements: Though images are captured within one second, motion of the subjects' faces can be source of severe data corruption. A comfortable seat with a headrest was used to diminish the subject movements during long acquisition sessions. However, this problem can also happen for instance due to breathing or muscle contractions during expressions. Therefore, faces that were deemed to be seriously faulty were re-captured. In the database, movement noise emerges especially in case of expressions, but depends on the subject and occasionally occurs. An example is shown at the middle row of Fig. 3.

Hairs and Eyes: Data on hair and facial hair, such as beard and eyebrows, generally causes spiky noise. Spiky surfaces arise also over the eyes. Basic smoothing filtering reduces these types of noises (Fig. 3).

Self-occlusions: Since data are captured from single views with this system, self-occlusions occur. The consequences are holes in the facial data, and uncompleted and distorted facial contours. Holes are formed due to missing data, mostly at the sides of the nose. Even slight head rotations generate high amount of self occlusions. In Fig. 3 these problems can be observed. Any processing was not performed for these problems.

Discontinuity: Discontinuity problems develop either inside the mouth when mouth is open, or in occluded face scans. The reconstruction of depth values at these discontinuous regions can sometimes be faulty. These errors are corrected by manual intervention using the system's software (Fig. 3).

5. Conclusion and Future Work

We have described the components, merits and limitations of a 3D face database, rich in Action Units, emotional expressions, head poses and types of occlusions. The involvement of actors/actresses, especially in the case of expressions, is considered to be an advantage.

We are planning several avenues of research on this database. Face recognition experiments have already been carried out on this database. These experiments consider the effect of face registration on the identification performance when the reference face model is obtained from neutral faces while test faces contain a variety of expressions. This research is presented in a companion paper [12]. Another research path is that of automatic facial landmarking. Automatically located landmarks can be used as initial steps for better registration of faces, for expression analysis and for animation. Various algorithms ranging from active appearance models to bunch graphs and statistical matched filter are studied.

For facial analysis and synthesis applications, non-rigid registration of faces is a very important intermediate step. Although variations due to expressions can be analyzed by rigid registration or landmark-based non-rigid registration methods, more faithful analysis can only be obtained with detailed non-rigid registration. Improved registration with non-rigid methods facilitates automatic expression understanding, face recognition under expressions and realistic face synthesis studies. Non-rigid registration is quite an ill-posed problem and needs further attention. We are conducting research along this line as well.

Acknowledgements. We would like to thank to subjects who voluntarily let their faces to be scanned. We are very thankful to Aydın Akyol from Istanbul Technical University; Jana Trojanova from West Bohemia University; Semih Esenlik and Cem Demirkır from Boğaziçi University; Nesli Bozkurt, İlkay Ulusoy, Erdem Akagündüz and Kerem Çalışkan from Middle East Technical University; and T. Metin Sezgin from University of Cambridge. They worked and helped for the collection and processing of this database during the Enterface'07 workshop project [13]. This work is supported by TÜBİTAK 104E080 and 103E038 grants.

References

1. P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, 1978
2. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Volume 1. (2005) 947–954 vol. 1
3. Lijun Yin; Xiaozhou Wei; Yi Sun; Jun Wang; Matthew J. Rosato, A 3D Facial Expression Database For Facial Behavior Research, 7th Int. Conference on Automatic Face and Gesture Recognition (FGR06), 10-12 April 2006 P:211 – 216
4. Faltemier, T.C., Bowyer, K.W., Flynn, P.J.: Using a multi-instance enrollment representation to improve 3d face recognition. In: Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE Int. Conf. (2007) 1–6
5. Heseltine, T., Pears, N., Austin, J.: Three-dimensional face recognition using combinations of surface feature map subspace components. Image and Vision Computing 26 (March 2008) 382–396
6. Zhong, C., Sun, Z., Tan, T.: Robust 3d face recognition using learned visual codebook. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. (2007) 1–6
7. Moreno, A., Sánchez, A.: Gavabdb: A 3d face database. In: Proc. 2nd COST275 Workshop on Biometrics on the Internet. (2004)
8. Beumier, C., Acheroy, M.: Face verification from 3d and grey level cues. Pattern Recognition Letters 22 (2001) 1321–1329
9. P. Ekman and W. V. Friesen, Constants Across Cultures in the Face and Emotion, Journal of Personality and Social Psychology, 17(2):124-129, 1971
10. Wallraven, C., D.W. Cunningham, M. Breidt and H.H. Bühlhoff: View dependence of complex versus simple facial motions. Proceedings of the First Symposium on Applied Perception in Graphics and Visualization, 181. (Eds.) Bühlhoff, H. H. and H. Rushmeier, ACM SIGGRAPH (2004)
11. <http://www.inspeck.com/>
12. Alyüz, N., Gökberk, B., Dibeklioglu, H., Savran, A., Salah, A. A., Akarun, L., Sankur, B., 3D Face Recognition Benchmarks on the Bosphorus Database with Focus on Facial Expressions. First European Workshop on Biometrics and Identity Management Workshop(BioID 2008)
13. Savran, A., Celiktutan, O., Akyol, A., Trojanova, J., Dibeklioglu, H., Esenlik, S., Bozkurt, N., Demirkir, C., Akagunduz, E., Caliskan, K., Alyuz, N., Sankur, B., Ulusoy, I., Akarun, L., Sezgin, T.M.: 3D face recognition performance under adversarial conditions. In: Proc. eINTERFACE07 Workshop on Multimodal Interfaces. (2007)

3D Face Recognition Benchmarks on the Bosphorus Database with Focus on Facial Expressions

Neşe Alyüz¹, Berk Gökberk², Hamdi Dibeklioglu¹, Arman Savran³, Albert Ali Salah⁴, Lale Akarun¹, Bülent Sankur³

¹ Boğaziçi University, Computer Engineering Department
nese.alyuz,hamdi.dibeklioglu,akarun@boun.edu.tr

² Philips Research, Eindhoven, The Netherlands
berk.gokberk@philips.com

³ Boğaziçi University, Department of Electrical and Electronics Engineering
arman.savran,bulent.sankur@boun.edu.tr

⁴ Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands
a.a.salah@cw.nl

Abstract. This paper presents an evaluation of several 3D face recognizers on the Bosphorus database which was gathered for studies on expression and pose invariant face analysis. We provide identification results of three 3D face recognition algorithms, namely generic face template based ICP approach, one-to-all ICP approach, and depth image-based Principal Component Analysis (PCA) method. All of these techniques treat faces globally and are usually accepted as baseline approaches. In addition, 2D texture classifiers are also incorporated in a fusion setting. Experimental results reveal that even though global shape classifiers achieve almost perfect identification in neutral-to-neutral comparisons, they are sub-optimal under extreme expression variations. We show that it is possible to boost the identification accuracy by focusing on the rigid facial regions and by fusing complementary information coming from shape and texture modalities.

1 Introduction

3D human face analysis has gained importance as a research topic due to recent technological advances in 3D acquisition systems. With the availability of affordable 3D sensors, it is now possible to use three-dimensional face information in many areas such as biometrics, human-computer interaction and medical analysis. Especially, for automatic face recognition, expression understanding, and face/facial feature localization problems, three-dimensional facial data offers better alternatives over using 2D texture information alone [1]. The information loss when projecting the inherently 3D facial structure to a 2D image plane is the major factor that complicates the task of analyzing human faces. Problems arise especially when adverse situations such as head pose variations, changes in illumination conditions, or extreme facial expressions are present in the acquired

data. The initial motivation for the exploitation of 3D information was to overcome these problems in human facial analysis. However, most of the proposed solutions are still limited to controlled acquisition conditions and constrained to frontal and mostly neutral 3D faces. Although there are increasing number of studies that focus on pose and/or expression invariant face recognition, the databases upon which they are based have not been systematically constructed for the analysis of these variations or they remain limited in scope. For example, the most frequently used 3D face database, the Face Recognition Grand Challenge (FRGC) database [2], contains mostly frontal faces with slight arbitrary pose variations. In the FRGC database, there are several acquisitions for different expressions which are labeled according to the emotions such as sadness and happiness. Comparison of publicly available 3D face databases in terms of pose, expression and occlusion variations can be found in [3].

The desiderata of a 3D face database enabling a range of facial analysis tasks ranging from expression analysis to 3D recognition are the following: i) Action units (FACS) [4], both single and compound; ii) Emotional expressions; iii) Ground-truthed poses; iv) Occlusions originating from hair tassel and a gesticulating hand. Motivated by these exigencies, we set out to construct a multi-attribute database. In this paper, we present the characteristics of the database collected as well as preliminary results on face registration and recognition.

2 The Bosphorus 3D Face Database

The Bosphorus database is a multi-expression, multi-pose 3D face database enriched with realistic occlusions such as hair tassel, gesticulating hand and eye-glasses [5, 3]. The variety of expressions, poses and occlusions enables one to set up arbitrarily challenging test situations along the recognition axis or along the expression analysis axis. We want to point out the opportunities that the Bosphorus database provides for expression understanding. The Bosphorus database contains two different types of facial expressions: 1) expressions that are based on facial *action units* (AU) of the Facial Action Coding System (FACS) and 2) *emotional expressions* that are typically encountered in real life. In the first type, a subset of action units are selected. These action units are grouped into three sets: i) 20 lower face AUs, ii) five upper face AUs and iii) three AU combinations. In the second type, we consider the following six universal emotions: happiness, surprise, fear, sadness, anger and disgust. Figure 1(b) shows all different types of expressions. To the best of our knowledge, this is the first database where ground-truthed action units are available. In order to achieve more natural looking expressions, we have employed professional actors and actresses.

Facial data are acquired using Inspeck Mega Capturor II 3D, which is a commercial structured-light based 3D digitizer device [6]. The 3D sensor has about $x = 0.3mm$, $y = 0.3mm$ and $z = 0.4mm$ sensitivity in all dimensions and a typical pre-processed scan consists of approximately 35K points. The texture images are high resolution (1600×1200) with perfect illumination conditions. The locations of several fiducial points are determined manually on both 2D

and 3D images. On each face scan, 24 points are marked on the texture images provided that they are visible in the given scan. The landmark points are shown in Figure 1(a).

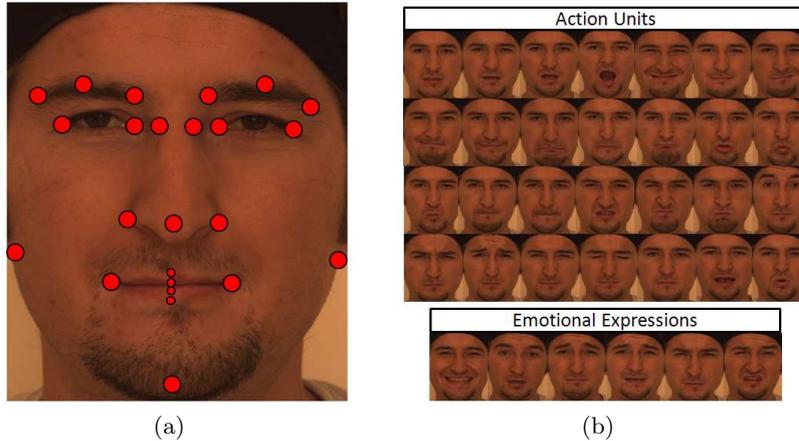


Fig. 1. a) Manually located landmark points and b) expressions for the Bosphorus database.

The Bosphorus database contains 3396 facial scans of 81 subjects. There are 51 men and 30 women in the database. Majority of the subjects are Caucasian and aged between 25 and 35. The Bosphorus database has two parts: the first part, Bosphorus v.1, contains 34 subjects and each of these subjects has 31 scans: 10 types of expressions, 13 different poses, four occlusions, and four neutral/frontal scans. The second part, Bosphorus v.2, has more expression variations. In Bosphorus v.2, there are 47 subjects having 53 scans¹. Each subject has 34 scans for different expressions, 13 scans for pose variations, four occlusions and one or two frontal/neutral face. 30 of these 47 subjects are professional actors/actresses.

3 Face Recognition Methodology

In this work, we apply commonly used techniques in face recognition to provide benchmarks for further studies. We have selected five face recognition approaches: three of them use shape information, and two use facial texture information. Two of the shape-based approaches are based on the Iterative Closest Point (ICP) algorithm, namely *one-to-all ICP* and average face model-based ICP (*AFM-based ICP*). The third one employs PCA coefficients obtained from 2D depth images. These techniques are explained in detail in Section 3.2. Texture-based approaches use either raw pixel information or PCA coefficients (eigenface

¹ Some subjects have fewer than 53 scans due to acquisition errors

technique). Before proceeding to identification methods, it is worthwhile to mention landmarking of faces because all these methods heavily rely on the quality of the initial alignment of facial surfaces.

3.1 Landmarking

Almost all 3D face recognition algorithms first need an accurate alignment between compared faces. There are various methods to align faces and most of them require several landmark locations that are easily and reliably detectable. ICP-based approaches which are explained later in this section, usually require these points at the initialization step. In our work, in addition to using 22 manually located landmark coordinates, we employ an automatic landmark localization method which estimates these points using the shape channel. The automatic landmarking algorithm consists of two phases [7]. In the first phase, a statistical generative model is used to describe patches around each landmark. During automatic localization, patches extracted from the facial surface are analyzed with these statistical models, and the region that produces the best likelihood value for each corresponding model is selected as the location of a landmark. A coarse-to-fine strategy is used to keep the search fast. We use inner and outer eye corners, nose tip and mouth corners, as these landmarks correspond to discriminative local structures. Figure 2(a) and 2(b) in Section 4 shows automatically found landmarks for a sample face image.

3.2 Shape-based Matchers

One-to-All ICP Algorithm: The 3D face recognition problem can be considered as a special case of a 3D object recognition problem. The similarity between two objects is inferred by features calculated from 3D models. Notice that most approaches require precise alignment (registration) of objects before similarity calculation and the performance depends heavily upon the success of registration [1].

The Iterative Closest Point (ICP) algorithm [8] has been one of the most popular registration techniques for 3D face recognition systems due to its simplicity. The ICP algorithm basically finds the best rigid transformation (i.e., translation, scale, and rotation matrices) to align surface A to surface B. Traditionally, a probe face is registered to *every* gallery face and an estimate of the volumetric difference between aligned facial surfaces is used as a dissimilarity measure. Therefore, we call this method *one-to-all ICP*. If we assume 3D point cloud representations of faces, dissimilarity can be estimated by the sum of the distances between corresponding point pairs in given facial surface pair. Indeed, ICP uses this measure during its iterations and after convergence, it outputs this dissimilarity measure as the alignment error.

AFM-based ICP Algorithm: The one-to-all ICP approach requires as many alignments as the size of the gallery set, this easily becomes infeasible when

the gallery set size is large. **For example, NESEDEN SURE BILGISI GELICEK.** An alternative approach would be to use a generic face model. All gallery faces are registered to this generic face model offline, before the identification phase [9], [10]. Thereby, only alignments between the probe faces and the generic face are needed to compute dissimilarities for the whole gallery set. This approach significantly shortens the identification delay by reducing the time complexity of the alignment phase. In the rest of the paper, we refer to this method as *AFM-based registration*.

Depth Image-based PCA Algorithm Most 3D sensors provide shape data in the form of 3D point clouds for the visible part of the object being scanned. For frontal facial 3D scans, the visible region usually contains the ear-to-ear frontal part of a human face. Therefore, there is at most one depth measurement, i.e., z coordinate, for any (x,y) coordinate pair. Due to this property, it is possible to project 2.5D data to an image plane where the pixels denote depth values. Images constructed in this way are called *depth images* or *range images*. 3D data should undergo post-processing stages during the conversion to depth images. Surface fitting is one of the important post-processing steps. A practical option for surface fitting is to obtain 3D triangulation of point cloud data and then to estimate the surface points inside the triangular patches by bilinear interpolation. Except for steep regions, such as the sides of the nose, information loss is minimal in depth image construction. Once 3D information is converted to 2D images, numerous approaches employed for 2D texture-based face recognition systems can be used for 3D face identification. Among them, using PCA coefficients as features is usually accepted as a baseline system for 3D depth image-based recognition. In our work, we perform whitening after computing PCA coefficients and use cosine distance for similarity calculation. As a pattern classifier, 1-nearest neighbor algorithm produces the estimated class label.

3.3 2D Texture Matchers

The Bosphorus database contains high quality texture information for each 3D facial model. In order to compare the performances of shape and texture channels we also implemented two 2D recognizers. The first, pixel-based method, simply uses gray-scale pixel information to represent a face. Texture images are normalized by scaling with respect to eye-to-eye distances. Illumination variations are handled by histogram equalization. In the pixel-based method, we use two regions: i) the whole face and ii) the upper facial region to test expression sensitivity. The second texture-based approach is the Eigenface technique where each face is transformed to a subspace by a PCA projection. As in the depth image method, we perform whitening and use 1-nearest neighbor classifier.

4 Experimental Results

We have performed recognition experiments on a subset of the Bosphorus database. The selected subset contains only neutral and expression-bearing images without

any pose variations or occlusions. Only one neutral image per person is used for enrollment, and the rest are used as the test set. First three rows of Table 1 show three experimental configurations. For the Bosphorus v.1, we have two experiments: one with the neutral probe set and the other with the non-neutral probe set. For v.2, there is only one experiment containing all non-neutral images of every subject in the probe set.

We have analyzed the effect of the number of landmarks and the effect of automatically detected landmarks in our tests. We use several subsets of landmarks that are presented in Figure 1(a). The performance of the automatic landmark detection module is summarized in Figure 2(c). We see that the most successful landmarks are the inner eye corners. In approximately 80% of the cases, they are found within tolerance, where the tolerance threshold is defined as 10% of the eye-to-eye distance. In general, inner eye corners and nose tip can be detected successfully, but outer eyebrows, and chin tip point usually can not be localized efficiently. The performance of the depth-image based automatic landmark detection is low. However, we include it here to test the performance of face recognizers with automatic landmarks.

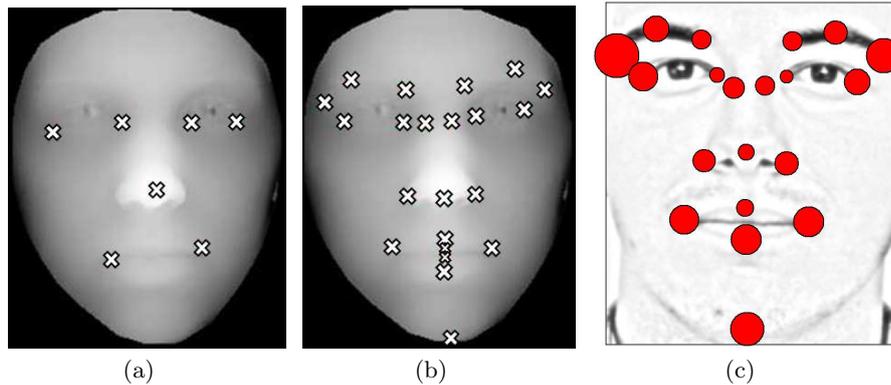


Fig. 2. Automatically located landmarks: the locations of a) seven fiducial landmarks found by the first phase, b) all 22 landmarks after the second phase, and c) the performance of automatic landmarking. Circle size denotes average pixel distance error for each landmark location.

We have performed recognition experiments on the v.1 and v.2 expression subsets, as summarized in Table 1. The first experiment was the one-to-all ICP experiment (One-to-All ICP_{M22} method in Table 1): Although this takes a long time, we provide these results as a benchmark. In the ICP coarse alignment stage, we used the 22 manually detected landmarks. As observed in Table 1, one-to-all ICP yields 99.02% correct identification on the v.1 neutrals. However, the performance drops to 74.04% for v.1 non-neutral and to 72.41% for v.2. This performance drop is to be expected, since the gallery includes only one neutral face. Next, we compare the AFM approach with the one-to-all ICP. The

AFM approach is very fast since it performs only one match. The results of this approach with 22 manually detected landmarks is denoted as AFM_{M22} in Table 1. On the v.1 database, AFM based identification classifies every facial image in the neutral probe accurately. However, in the non-neutral v.1 probe set, the correct classification rate drops to 71.39%. For v.2 tests, only 67.67% of the probe set is identified correctly. On comparison with one-to-all results, we see that AFM performs better on neutral faces, but suffers a small drop in performance in faces with expressions. Since this drop is not very large, we use the AFM approach for the rest of the tests.

Table 1. Correct classification rates (%) of various methods on the Bosphorus database. Coarse alignment configurations used in these methods are denoted as subscripts: M and A is for manual and automatic landmarking, respectively. The numbers used in the subscripts denote the number of landmarks used; i.e., AFM_{M5} is the AFM method aligned with five manual landmark points.

Method	v.1	v.1	v.2
	Neutral	Non-neutral	Non-neutral
Gallery Set Size	34	34	47
Probe Set Size	102	339	1508
AFM_{M5}	99.02	69.62	65.12
AFM_{M7}	100.00	73.75	68.83
AFM_{M8}	99.02	72.27	69.36
AFM_{M22}	100.00	71.39	67.67
AFM_{A7}	80.39	62.24	-
AFM_{A22}	81.37	62.24	-
One-to-All ICP $_{M22}$	99.02	74.04	72.41
DI – PCA $_{M22}$ (Whole face)	100.00	71.09	70.56
DI – PCA $_{M22}$ (Eye,Nose)	100.00	85.55	88.79
TEX-Pixel (Whole face)	97.06	93.51	92.52
TEX-Pixel (Upper face)	97.06	90.56	92.59
TEX-Eigenface (Whole Face)	97.06	87.61	89.25
Fusion of AFM_{M7} and TEX-Pixel (Whole face)	-	-	95.09
Fusion of DI – PCA $_{M22}$ (Eye,Nose) and and TEX-Pixel (Whole face)	-	-	98.01

The effect of facial landmarks on the identification rate is next analyzed. For this purpose, we look further into two quantities: 1) The subset of facial landmarks that should be used in coarse alignment and 2) The performance change caused by the use of automatic landmark localizer. For the first case, we formed three landmark subsets of size five, seven and eight. Landmark subset of size five only uses landmark points around the nose. The landmark set with seven landmarks contains eye corner points, nose tip, and mouth corners. The eight-point subset is the same as the seven-point set but with the added chin tip point. We see that using only seven landmarks leads to better performance than using all 22 landmarks. Accuracy in v.1 non-neutral set is 73.75% (see

Table 1, marked as AFM_{M7}) and in v.2, it is 68.83%. If faces are registered according to the nose region only (using five landmarks, AFM_{M5} in Table 1), we see degradation in accuracy. Adding chin tip to the previously selected seven landmarks does not change the identification rate significantly.

If we turn back to our second question about the effect of automatic landmarking on the identification rates, we see significant performance drop with automatic landmarking. Entries marked as AFM_{A7} and AFM_{A22} in Table 1 show that, irrespective of which landmark subset is used, there is approximately 20% and 10% accuracy decrease in neutral and non-neutral probe sets, respectively. This is mostly due to the localization errors in landmark detection.

Regarding all ICP-based experiments, we see that AFM_{M7} presents a good compromise in that: i) It is computationally much faster than one-to-all performance and performs only a little worse; and ii) It relies on only 7 landmarks, which are easier to find.

The next set of experiments are with the depth image PCA method ($DI - PCA_{M22}$ methods in Table 1). We have tried two versions: Using the whole face, and using only the eyes and the nose regions. Both perform perfectly with the neutral faces in v.1. In non-neutral v.1, and v.2, the performance of the whole face is 71.09% and 70.56%, respectively. When only the eye and nose regions are included, performance rises to 85.55% in v.1 non-neutrals and to 88.79% for v.2. Overall, we see that local PCA-based representation of eye and nose region is the best shape modality-based recognizer.

We have also used 2D textures to classify the faces. We have obtained very good identification performance with texture images. Note that the texture images are of very high quality, with perfect illumination and high resolution. The performance obtained with texture pixels is reported for i) the whole face and ii) the upper part (denoted as $TEX\text{-Pixel}$ in Table 1). The eigenface technique is also applied ($TEX\text{-Eigenface}$). Identification performances of all three algorithms on the neutral v.1 are identical: 97.06%. On the non-neutral v.1, the three algorithms obtain 93.51%, 90.56%, and 89.25%, respectively. Recognition performance on the v.2 are unexpectedly higher: 92.52%, 92.59% and 89.25%, respectively. We note that the texture performances are higher than the shape performances. This is due to the perfect illumination conditions and the high resolution of the 2D images

And lastly, we fuse the results of the 3D and 2D classifiers. Using product rule to combine the dissimilarity scores of AFM-based ICP method and pixel-based textural classifier (See Table 1, Fusion of AFM_{M7} and $TEX\text{-Pixel}$), we achieve 95.09% correct identification rate in the v.2 experiment. If $DI\text{-PCA}$ of the eye/nose region is used as a shape classifier in fusion, 98.01% accuracy is obtained (See Figure 3(b) for all 30 images misclassified in the v.2 set). Cumulative matching characteristic (CMC) curves of local $DI\text{-PCA}$ and texture classifiers, together with their fusion performance, are shown in Figure 3(a). Notice that although rank-1 performance of the texture classifier is higher, shape classifier becomes superior after rank 3.

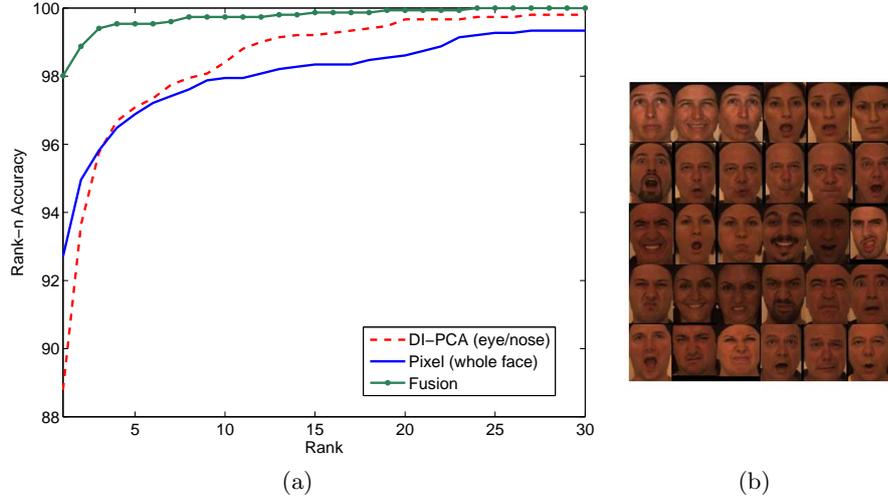


Fig. 3. a) CMC curve for i) local PCA based depth image algorithm, ii) pixel-based texture algorithm and iii) their fusion, and b) Misclassified faces in the v.2 set by the fusion of DI-PCA and TEX-Pixel method.

5 Conclusion

In this work, benchmarking studies on a new challenging 3D face database are presented. We have used 3D recognition methods with proven performance: Two of these algorithms use ICP alignment for dissimilarity calculation. One is based on generic face template (AFM) for fast registration, and the other exhaustively searches the closest face model from the gallery set for a given probe image. In addition to ICP-based methods, depth images are also used where feature construction is handled via the PCA technique.

3D cameras almost always yield 2D texture images in addition to 3D data. At close range and under good illumination, the texture images turn out to be of high quality. In fact, texture images singly or in complementary role to 3D data can boost the performance. In our study, fusion of the shape and texture based methods has yielded recognition performances as high as 98.01%. The main conclusions of our work are as follows:

- The performance obtained with the one-to-all registration is comparable to that of AFM registration, both with neutral and expression faces. On the other hand, AFM method is orders of magnitude faster. Therefore AFM is preferable.
- The 3D recognition performance suffers heavily from inexactitude of landmarks. The present landmarking algorithm causes a heavy performance drop of 10-20% percentage points. Therefore real-time and reliable face landmarking remains still an open problem.

- Depth images with PCA form a viable competitor to the 3D point cloud feature set, and in fact outperform it. It remains to see if alternative feature sets, e.g., subspace methods or surface normals can bring improvements.
- The fusion of 2D texture and 3D shape information is presently the scheme with the highest performance.

The Bosphorus database is suitable for studies on 3D human face analysis under challenging situations such as in the presence of occlusion, facial expression, pose variations. The future work will consist of i) improving landmark localization performance, ii) testing the sensitivity of 3D face recognition algorithms under pose changes, and iii) employing different representation methods other than point clouds and depth images.

6 Acknowledgements

This work is supported by TÜBİTAK 104E080 and 103E038 grants.

References

1. Bowyer, K., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer Vision and Image Understanding* **101** (2006) 1–15
2. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: *Proc. of. Computer Vision and Pattern Recognition*. Volume 1. (2005) 947–954
3. Savran, A., Alyüz, N., Dibekliöglü, H., Çeliktutan, O., Gökberk, B., Akarun, L., Sankur, B.: Bosphorus database for 3D face analysis. In: *Submitted to the First European Workshop on Biometrics and Identity Management Workshop(BioID 2008)*
4. Ekman, P., Friesen, W.: *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press (1978)
5. Savran, A., Çeliktutan, O., Akyol, A., Trojanova, J., Dibekliöglü, H., Esenlik, S., Bozkurt, N., Demirkır, C., Akagündüz, E., Çalıskan, K., Alyüz, N., Sankur, B., Ulusoy, İ., Akarun, L., Sezgin, T.M.: 3D face recognition performance under adversarial conditions. In: *Proc. eNTERFACE07 Workshop on Multimodal Interfaces*. (2007)
6. Inspeck Mega Capturor II Digitizer: <http://www.inspeck.com/>
7. Salah, A.A., Akarun, L.: 3D facial feature localization for registration. In: *Proc. Int. Workshop on Multimedia Content Representation, Classification and Security LNCS*. Volume 4105/2006. (2006) 338–345
8. Besl, P., McKay, N.: A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(2) (1992) 239–256
9. İrfanoğlu, M., Gökberk, B., Akarun, L.: 3D shape-based face recognition using automatically registered facial surfaces. In: *Proc. ICPR*. Volume 4. (2004) 183–186
10. Gökberk, B., İrfanoğlu, M., Akarun, L.: 3D shape-based face representation and feature extraction for face recognition. *Image and Vision Computing* **24**(8) (2006) 857–869

Identity Management in Face Recognition Systems

Massimo Tistarelli and Enrico Grosso

Università di Sassari – Computer Vision Laboratory,
Alghero, Italy
{tista,grosso}@uniss.it

Abstract. Face recognition is one of the most challenging biometric modalities for personal identification. This is due to a number of factors, including the complexity and variability of the signal captured by a face device. Several issues incur in the management of a face template as user's identity. Data dimensionality reduction, compactness of the representation, uniqueness of the template and ageing effects, are just but a few of the issues to be addressed. In this paper we present the current state of the art in face recognition technology and how this related to the proper management of a user's identity. Some real cases are presented and some conclusions are drawn.

Keywords: face recognition, biometrics, identity management, pattern recognition, computer vision.

1 Introduction

A recent poll from Harris Interactive, involving 1,000 adults within US, shows that the majority of US citizens would favor an increase in surveillance systems to increase security. 70% of the interviewed were in good favor of expanded camera surveillance on streets and in public places. This is but one of the many recalls to the impact of biometric research in social life. The increasing need for reliable security systems, in turn, highlights the need for pursuing advanced research in the field. It is not the case, not anymore, that a simple algorithmic solution can provide the answer to the emerging needs. It is rather important that reliable, easy to use, and smart systems are devised and introduced in the society.

Security is primarily advocated in recurrent scenarios such as, street surveillance and access control. In these applications recognition at a distance is the key element for a successful identification. There are not as many viable solution for identification at a distance. Even though several remarkable examples are emerging from iris and gate technologies, today's most reliable systems are those based on face recognition.

Face recognition/verification has attracted the attention of researchers for more than two decades and it is among the most popular research areas in the field of computer vision and pattern recognition. Several approaches have been proposed for face recognition based on 2D and 3D images. In general, face recognition technologies are based on a two step approach:

- an off-line enrollment procedure is established to build a unique template for each registered user. The procedure is based on the acquisition of a pre-defined set of face images (or a complete video), selected from the input image stream, and the template is build upon a set of features extracted from the image ensemble;
- an on-line identification or verification procedure where a set of images are acquired and processed to extract a given set of features. From these features a face description is built to be matched against the user's template.

Regardless of the acquisition devices exploited to grab the image streams, a simple taxonomy can be based on the computational architecture applied to: extract powerful features for identification and to derive a template description for subsequent matching. The two main algorithmic categories can be defined on the basis of the relation between the subject and the face model, i.e. whether the algorithm is based on a subject-centered (eco-centric) representation or on a camera-centered (ego-centric) representation. The former class of algorithms relies on a more complex model of the face, which is generally 3D or 2.5D, and it is strongly linked with the 3D structure of the face.

These methods rely on a more complex procedure to extract the features and build the face model, but they have the advantage of being intrinsically pose-invariant. The most popular face-centered algorithms are those based on 3D face data acquisition and on face depth maps. The ego-centric class of algorithms strongly relies on the information content of the gray level structures of the images. Therefore, the face representation is strongly pose-variant and the model is rigidly linked to the face appearance, rather than to the 3D face structure. The most popular image-centered algorithms are the holistic or subspace-based methods, the feature-based methods and the hybrid methods.

Over these fundamental classes of algorithms several elaborations have been proposed. Among them, the kernel methods greatly enhanced the discrimination power of several ego-centric algorithms, while new feature analysis techniques such as the local binary pattern (LBP) representation greatly improved the speed and robustness of Gabor-filtering based methods. The same considerations are valid for eco-centric algorithms, where new shape descriptors and 3D parametric models, including the fusion of shape information with the 2D face texture, considerably enhanced the accuracy of existing methods.

2 Face biometric technologies

Gartner Group in 2005 recognized biometrics to be one of the most promising IT technologies for the future. The graph in figure 1 well represents the expected follow-up of biometrics in the IT research and market for the near future. AT the same time, biometric apparently received more attention from the media and advertising companies, than the real application breakthrough. This, in turn, indicates the requirement for an increased focus on killer applications and a closer involvement of industries in research.

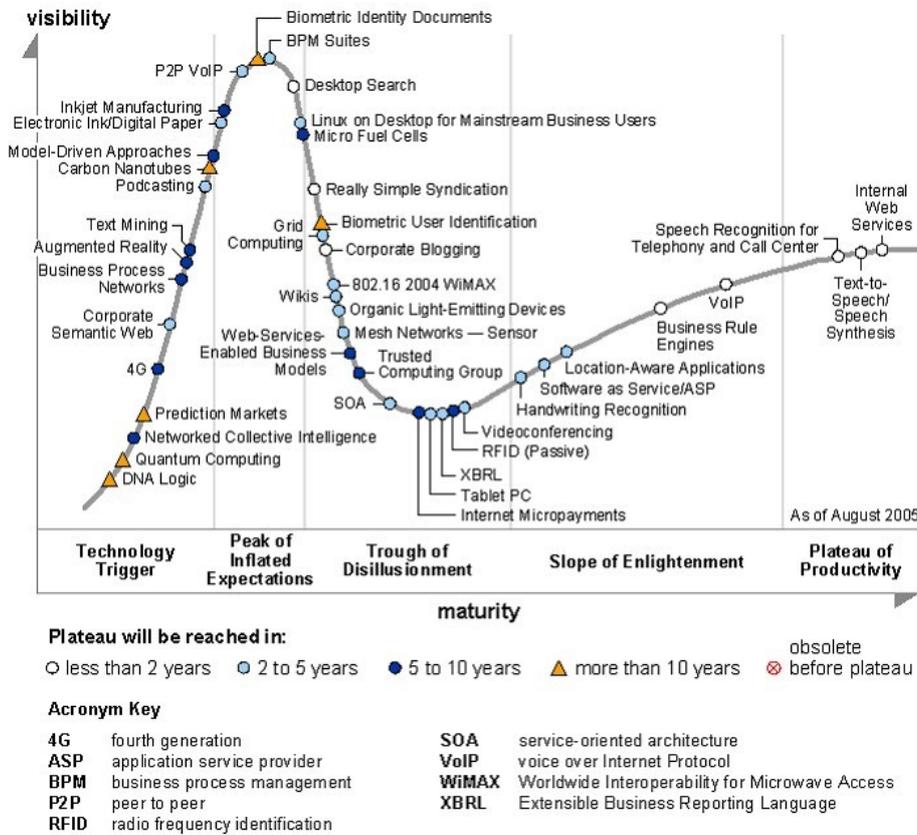


Fig. 1. Gartner’s group graph representing the technology trend in IT.

While several industrial products for deploying face recognition already exist, still there is a great need to enhance the basic technologies implied. For example, ageing, spoofing and illumination compensation are still open issues which require to be addressed. At the same time, the proper management of the user’s identity can not be viewed as detached from the algorithms applied to process the raw signal and to build the biometric template. In the case of face biometrics several techniques have been proposed which can be broadly divided into two main categories: image-centered (eco-centric) and subject-centered (ego-centric). In this paper the eco-centric methods will be considered as well as their impact in the management of the user’s identity.

Almost all biometric identification techniques, including face-based methods, rely on a two step process. In the first step a set of features are extracted from the images. In the second step the extracted features are fed into a classifier to actually identify the class to which the probe face belongs. The classification is a crucial process which can be easily tailored to any feature representation. Once the classifier is tuned to the adopted feature representation, it must be also tuned to the population of subjects to be correctly classified. Toward this end it is necessary to model each possible instance

of all possible classes (the “rest of the world”) to define the discrimination parameters (the classification threshold) to distinguish the representation of one subject from every other subject.

On the contrary, the feature extraction heavily depends on the template representation and on the physical characteristics of the raw biometric signal. This leads to a variety of different (and incompatible) methods to draw face-based identity representations. The differences include the template size, the discrimination power and the compactness of the representation.

The heterogeneity in the face-based representations led to a proliferation of identification systems and industrial solutions which are hardly compatible. The remainder of the paper tries to summarize the main advantages and drawbacks of the main algorithms for face-based identity representation and matching.

2.1 Subspace methods

The most popular techniques for frontal face identification and authentication are the subspace methods. These algorithms consider the entire image as a feature vector with the aim to find projections (bases) that optimize some criterion defined over the feature vectors that correspond to different classes. Then the original high dimensional image space is projected into a low dimensional one. In all these approaches the face representation (the *template*) is a vector of eigenvalues determining the position of the subject's sample in the feature space, defined by the basis vectors. The classification is usually performed according to a simple distance measure in the final multidimensional space. The recurrent obsession in subspace methods is to reduce the dimensionality of the search space. A large database with 1000 gray level images with a resolution of 512x512 pixels, can be reduced to a small set of vectors, each with the same size of each sample image. Even though the dimensionality reduction is performed according to a minimization criterion to enhance some data features, the immediate effect of this process is to reduce the information content in the data. Dimensionality reduction always produces an effect similar to low-pass filtering, where the size of the data is reduced at the cost of a lower discrimination power.

Various criteria have been employed in order to find the bases of the low dimensional spaces. Some of them have been defined in order to find projections that they best express the population without using the information of how the data are separated to different classes. Another class of criteria is the one that deals directly with the discrimination between classes. Finally, statistical independence in the low dimensional feature space is a criterion that is used in order to find the linear projections.

The first method employed for low dimension representation of faces is the *eigenfaces* (PCA) approach [1]. This representation was used in [2] for face recognition. The idea behind the eigenface representation is to apply a linear transformation that maximizes the scatter of all projected samples. This operation corresponds to a singular value decomposition of the data ensemble. The PCA approach was extended to a nonlinear alternative using kernel functions (KPCA) [3]. Recently KPCA with fractional power polynomial kernel has been successfully used along with Gabor features for face recognition [4].

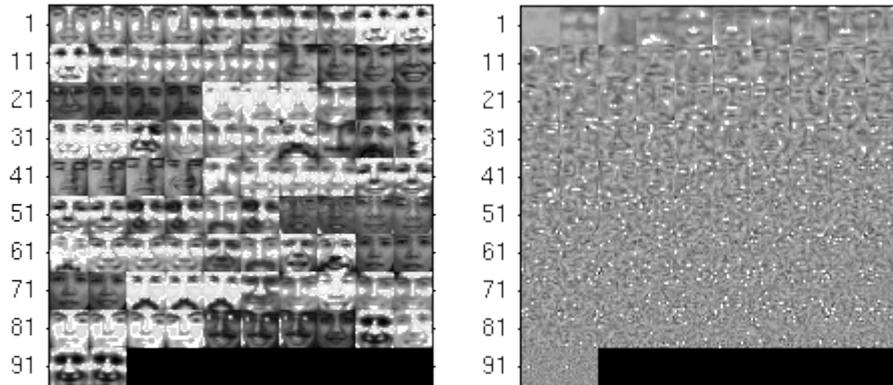


Fig. 2. (left) Sample database composed of 92 faces. (right) Set of 92 base vectors obtained with the PCA decomposition.

Another subspace method that aims at representing the facial face without using class information is the *non negative matrix factorization* (NMF) [5]. The NMF algorithm, like PCA, represents a face as a linear combination of bases. The difference with PCA is that it does not allow negative elements in both the bases vectors and the weights of the linear combination. This constraint results to radically different bases than PCA. On the one hand the bases of PCA are eigenfaces, some of which resemble distorted versions of the entire face. On the other hand the bases of NMF are localized features that correspond better to the intuitive notions of face parts [5]. An extension of NMF that gives even more localized bases by imposing additional locality constraints is the so-called *local non negative matrix factorization* (LNMF) [6].

Linear discriminant analysis (LDA) is an alternative method to PCA maximizing the separation among different classes (subjects). In [7,8], it was proposed to apply LDA in a reduced PCA space for facial image retrieval and recognition, the so-called *fisherfaces*. In this approach the PCA decomposition is first applied ensuring the scatter matrix to be non-singular. The dimension of the new features is further reduced by using *Fisher's Linear Discriminant* (FLD) optimization criterion to produce the final linear transformation. The drawback of this method is the low-pass effect produced by the PCA. The initial dimensionality reduction may sensibly reduce the discrimination power of the final representation [11].

To overcome this limitation, direct LDA (D-LDA) algorithms for discriminant feature extraction were proposed [9,10,11]. The DLDA algorithms are usually applied using direct diagonalization methods for finding the linear projections that optimize the discriminant criterion.

To make nonlinear problems tractable, LDA has been generalized to its kernel version, namely *general discriminant analysis* (GDA) [12] or *kernel Fisher discriminant analysis* (KFDA) [13]. In GDA the original input space is projected using a nonlinear mapping from the input space (the facial image space) to a high-dimensional feature space, where different classes of faces are supposed to be linearly separable. The idea behind GDA is to perform LDA in the feature space instead of the

input space. The interested reader can refer to [12-16.] for different versions of KFDA and GDA.

The main drawback of the methods that use discriminant criteria is that they may cause overtraining. Moreover, it is quite difficult to build a discriminant function on small training sample sets with reasonable generalization abilities [17,18]. This is true in many practical cases where a very limited number of facial images are available in database training sets. The small number of facial images, for each face class, affects both linear and the nonlinear methods where the distribution of the client class should be evaluated in a robust way [13]. In [19] it has been shown that LDA outperforms PCA only when large and representative training datasets are available.

In order to find linear projections that minimize the statistical dependence between its components the independent component analysis has been proposed [20, 21] for face recognition. ICA has been applied in the original input space of the facial images [20] or using Gabor based features of the facial images [21]. The nonlinear alternative of ICA using kernel methods has been also proposed in [22].

2.2 Elastic graph matching

The *elastic graph matching* (EGM) is a practical implementation of the *dynamic link architecture* (DLA) for object recognition [23]. In EGM, the reference object graph is created by overlaying a rectangular elastic sparse graph on the object image and calculating a Gabor wavelet bank response at each graph node. The graph matching process is implemented by a stochastic optimization of a cost function which takes into account both jet similarities and node deformation. A two stage coarse-to-fine optimization procedure suffices for the minimization of such a cost function.

In [24] it has been shown that EGM outperforms eigenfaces and self-associative neural networks for face recognition. In [25] the graph structure has been enhanced by introducing a stack like structure, the so-called *bunch graph*. In the bunch graph structure for every node a set of Gabor jets is computed for different instances of a face (e.g., with mouth open or closed, etc.). That way, the bunch graph representation covers a variety of possible face appearances [26]. Practical methods for increasing the robustness of EGM against translations, deformations and changes in background have been presented in [27,28].

Several variations of the standard EGM have been proposed [29-33]. Among them, is the *morphological elastic graph matching* (MEGM) where the Gabor features are replaced by multiscale morphological features, obtained through a dilation-erosion of the facial image [32]. In [29] the standard coarse to fine approach [28] for EGM is replaced by a simulated annealing method that optimizes a cost function of the jet similarity distances subject to node deformation constraints. The multiscale morphological analysis has given comparable verification results with the standard EGM approach, without the need to compute Gabor filter banks. Another variant of EGM has been presented in [33], where morphological signal decomposition has been used instead of the standard Gabor analysis.

To enhance the EGM performance, several techniques have been proposed weighting the graph nodes according to their relevance for recognition [29,33-35]. As

an example, linear discriminant techniques have been employed for selecting the most discriminating features [28,29,33,36,37].

In [34] the selection of the weighting coefficients was based on a nonlinear function that depends on a small set of parameters. These parameters have been determined on the training set by maximizing a criterion using the simplex method. In [29,33] the set of node weighting coefficient was not calculated by some criterion optimization but by using the first and second order statistics of the node similarity values. A Bayesian approach for determining which nodes are more reliable has been used in [26]. A more sophisticated scheme for weighting the nodes of the elastic graph, by constructing a modified class of support vector machines, has been proposed in [35]. It has been also shown that the verification performance of the EGM can be highly improved by proper node weighting strategies.

The subspace of the face verification and recognition algorithms consider the entire image as a feature vector and their aim is to find projections that optimize some criterion defined over the feature vectors that correspond to different classes. The main drawback of these methods is that they require the facial images to be perfectly aligned. That is, all the facial images should be aligned in order to have all the fiducial points (e.g. eyes, nose, mouth, etc.) represented at the same position inside the feature vector. For this purpose, the facial images are very often aligned manually and moreover they are anisotropically scaled. Perfect automatic alignment is in general a difficult task to be assessed. On the contrary, elastic graph matching does not require perfect alignment in order to perform well. The main drawback of the elastic graph matching is the time required for multiscale analysis of the facial image and for the matching procedure. A recent approach tries to overcome this limitation by using the *Shift Invariant Feature Transform* (SIFT) as graph nodes [38,39]. SIFT features can be extracted with a fast algorithm from the images, thus reducing the computation time required to build the representation. This enhanced method proved to produce superior performances than holistic methods, on standard databases.

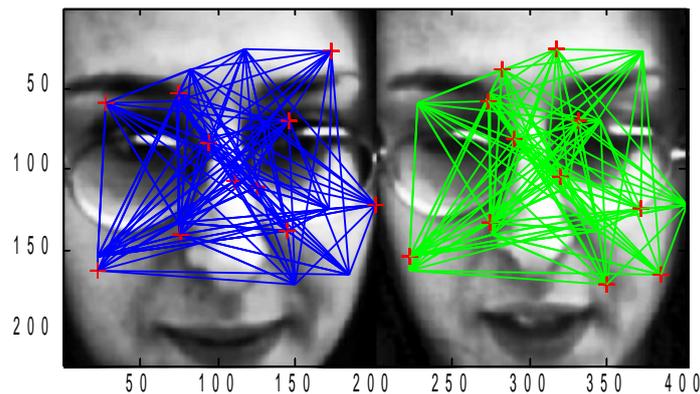


Fig. 3. Example graph constructed from a set of 11 SIFT feature points.

2.3 Dynamic face recognition

Historically face recognition and authentication has been treated as the matching between snapshots containing the representation of a face. In the human visual system the analysis of visual information is never restricted to a time-confined signal. Much information on the analysed visual data is contained within the temporal evolution of the data itself. Therefore a considerable amount of the “neural power” in humans is devoted to the analysis and interpretation of time variations of the visual signal.

On the other hand, processing single images considerably simplifies the recognition process. Therefore, the real challenge is to exploit the added information in the time variation of face images, limiting the added computational burden. An additional difficulty in experimenting dynamic face recognition is the dimensionality of the required test data. A statistically meaningful experimental test requires a considerable number of subjects (at least 80 to 100) with several views taken at different times. Collecting video streams of 4 to 5 seconds from each subject and for each acquisition session implies the storage and subsequent processing of a considerable amount (hundreds of Gigabytes) of data.

There are only few face recognition systems in the literature based on the analysis of image sequences. The developed algorithms generally exploit the following advantages from the video sequence:

1. The matching process is repeated over more images and the resulting scores are combined according to some criterion. Several approaches have been proposed to integrate multiple similarity measurements from video streams. Most of the proposed algorithms rely on the concept of data fusion [64] and uncertainty reduction [73].
2. The input sequence is filtered to extract the image data best suited for recognition. This method is often coupled with a template representation based on a sequence of face views. An example of this use is the IRDB (Incremental Refinement of Decision Boundaries) [81,89] where the face representation is dynamically augmented by processing and selecting subsequent frames in the input video stream on the basis of the output of a statistical classifier.
3. The motion in the sequence is used to infer the 3D structure of the face and perform 3D instead of 2D recognition [40]. An interesting similar approach is based on the generalization of classic single view matching to multiple views [40, 41] and the integration of video into a time-varying representation called “identity surfaces”.
4. Map the processing algorithm to extend the face template representation from 2D to 3D, where the third dimension is time. There are few examples of this approach including composite PCA, extended HMMs, parametric eigenspaces, multi-dimensional classifiers, neural networks and other, video oriented, integrated approaches.
5. Detect and identify facial expression either for face re-normalization or emotion understanding.

3 Face representations

3.1 Face representation from single images

Holistic methods for face identification require a large (statistically significant) training set to build the base vectors determining the low dimension space. The generalization capabilities of these methods have been tested to some extent but are still unclear. Up to now tests have been performed on databases with limited size. Even the FRGC database [93,94] only comprises few thousands subjects. Scaling up to larger databases, including hundred of thousands individuals, even if possible, would make the problem very difficult to be numerically analyzed. Managing the identity by these face representations requires to be able to discriminate each single individual through a single feature space, but this can be hardly guaranteed. The best performing face recognition methods, based on holistic processing, under real conditions reach an equal error rate (EER) around 1%. This corresponds to 100 wrongly classified subjects over a database of 10,000 individuals or 1,000 over 100,000. The template size depends on the dimensionality of the representation space i.e. the number of basis vectors selected for the database representation. This value depends on the population of subjects, the variability of the face appearance (pose, expression, lighting, etc.), the number of classes and the discrimination power to be achieved. Therefore, coping with many variations in the face appearance, for example to deal with ageing, the size of the subspace and hence the representation can become indefinitely large.

An advantage of EGM is the strong dependence on the input signal rather than on the population of subjects analyzed. Therefore, the subject's identity is represented exclusively from information related to data captured from each subject. The relation to the "rest of the world" is limited to the classification parameters which must be tuned for classification. The resulting EGM face template can be very compact as it is limited to the graph structure with the associated Gabor weights. This allows to cope with many variations, including ageing, without affecting the size of the representation.

The drawbacks of EGM stem from the adaptation of the graph to the subject's face. In order to be non-ambiguous it generally requires a good initialization.

3.2 Face representation from video streams

An advantage of processing face video over single images stems from the possibility to define "dynamic templates". These representations can exploit both physical and behavioral traits, thus enhancing the discrimination power of the classifier. The representation of the subject's identity can be arbitrarily rich at the cost of a large template size.

Several approaches have been proposed to generalize classical face representations based on a single-view to multiple view representations. Examples of this kind can be

found in [43,44] and [46-48] where face sequences are clustered using vector quantization into different views and subsequently fed to a statistical classifier.

Recently, Krüger, Zhou and Chellappa [49-57] proposed the “video-to-video” paradigm, where the whole sequence of faces, acquired during a given time interval, is associated to a class (identity). This concept implies the temporal analysis of the video sequence with dynamical models (e.g., Bayesian models), and the “condensation” of the tracking and recognition problems.

Other face recognition systems, based on the still-to-still and multiple stills-to-still paradigms, have been proposed [42,58,59]. However, none of them is able to effectively handle the large variability of critical parameters, like pose, lighting, scale, face expression, some kind of forgery in the subject appearance (e.g., the beard). Typically, a face recognition system is specialized on a certain type of face view (e.g. frontal views), disregarding the images that do not correspond to such view. Therefore, a powerful pose estimation algorithm is required.

In order to improve the performance and robustness, multiple classifier systems (MCSs) have been recently proposed [60].

Achermann and Bunke [61] proposed the fusion of three recognizers based on frontal and profile faces. The outcome of each expert, represented by a score, i.e., a level of confidence about the decision, is combined with simple fusion rules (majority voting, rank sum, Bayes’s combination rule). Lucas [43,44] used a n -tuple classifier for combining the decisions of experts based on sub-sampled images.

Other interesting approaches are based on the extension of conventional, parametric classifiers to improve the “face space” representation. Among them are the extended HMMs [72], the Pseudo-Hierarchical HMMs [91,92] and parametric eigenspaces [64], where the dynamic information in the video sequence is explicitly used to improve the face representation and, consequently, the discrimination power of the classifier. In [71] Lee *et al.* approximate face manifolds by a finite number of infinite extent subspaces and use temporal information to robustly estimate the operating part of the manifold.

There are fewer methods that recognize from manifolds without the associated ordering of face images. Two algorithms worth mentioning are the Mutual Subspace Method (MSM) of Yamaguchi *et al.* [83,90] and the Kullback-Leibler divergence based method of Shakhnarovich *et al.* [78]. In MSM, infinite extent linear subspaces are used to compactly characterize face sets i.e. the manifolds that they lie on. Two sets are then compared by computing the first three principal angles between corresponding principal component analysis (PCA) subspaces [48]. The major limitation of MSM is its simplistic modelling of manifolds of face variation. Their high nonlinearity invalidates the assumption that data is well described by a linear subspace. Moreover, MSM does not have a meaningful probabilistic interpretation.

The Kullback-Leibler divergence (KLD) based method [78] is founded on information-theoretic grounds. In the proposed framework, it is assumed that i -th person’s face patterns are distributed according to $p_i(x)$. Recognition is then performed by finding $p_j(x)$ that best explains the set of input samples – quantified by the Kullback-Leibler divergence. The key assumption in their work, that makes divergence computation tractable, is that face patterns are normally distributed.

4 Conclusions

The representation of faces strongly depends on the algorithm used to extract the facial features. Many techniques have been proposed to build face templates from still images and video. Even though holistic, subspace methods are the most widely used and studied, the template size can be very large if a large number of variation modes is required. On the other hand, the discrimination power of the algorithms strongly relies on the population adopted to perform the training of the subspace representation process. This, in turn, makes the subspace methods, strongly affected by the database. The same does not hold for other face representations, such as the elastic graph matching and other derived methods. On the other hand, while it can be relatively easy to adapt the representation to cope with the variability in the face appearance (for example due to ageing), the performances are greatly affected by the accuracy in the localization of the graph nodes on the face image.

Dynamic face representations are the most rich and compact at the same time. They can provide remarkable performances and a high discrimination power, at the cost of a larger template size. From the results provided in the literature this can be the right avenue to be pursued to build a robust and yet flexible identity representation.

References

1. M. Kirby and L. Sirovich: Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, Jan. 1990.
2. M. Turk and A. P. Pentland: Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
3. B. Schölkopf, A. Smola, and K. R. Müller: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, vol. 10, pp. 1299–1319, 1999.
4. L. Chengjun: Gabor-based kernel pca with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572 – 581, May 2004.
5. D.D. Lee and H.S. Seung: Learning the parts of objects by non-negative matrix factorization. *Nature*, vol. 401, pp. 788–791, 1999.
6. S.Z. Li, X.W. Hou, and H.J. Zhang: Learning spatially localized, parts-based representation. In *Proceedings CVPR*, 2001, pp. 207–212.
7. H. Yu and J. Yang: A direct lda algorithm for high-dimensional data with application to face recognition. *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.
8. L. Juwei, K.N. Plataniotis, and A.N. Venetsanopoulos: Face recognition using lda-based algorithms. *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.
9. G. Baudat and F. Anouar, “Generalized discriminant analysis using a kernel approach,” *Neural Comput.*, vol. 12, pp. 2385–2404, 2000.
10. K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf: An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 181-201, 2001.
11. L. Juwei, K.N. Plataniotis, and A.N. Venetsanopoulos: Face recognition using kernel direct discriminant analysis algorithms. *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 117–126, 2003.

- 12.S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R.Muller: Fisher Discriminant Analysis with Kernels. In Proceedings IEEE Int'l Workshop Neural Networks for Signal Processing IX, pp. 41-48, Aug. 1999.
- 13.S. Mika, G. Ratsch, B. Scholkopf, A. Smola, J. Weston, and K.-R. Muller: Invariant Feature Extraction and Classification in Kernel Spaces. Advances in Neural Information Processing Systems 12, Cambridge, Mass.: MIT Press, 1999.
- 14.A.K. Jain and B. Chandrasekaran: Dimensionality and sample size considerations in pattern recognition practice. In Handbook of Statistics, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam: North-Holland, vol. 2, pp. 835-855, 1987.
- 15.S.J. Raudys and A.K. Jain: Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 3, pp. 252-264, 1991.
- 16.A. Martinez and A. Kak, "PCA versus LDA," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 228-233, 2001.
- 17.M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski: Face recognition by independent component analysis. IEEE Transactions on Neural Networks, vol. 13, no. 6, pp. 1450-1464, 2002.
- 18.L. Chengjun and H. Wechsler: Independent component analysis of Gabor features for face recognition. IEEE Transactions on Neural Networks, vol. 14, no. 4, pp. 919-928, July 2003.
- 19.F. Bach and M.Jordan: Kernel Independent Component Analysis. Journal of Machine Learning Research, vol. 3, pp. 1-48, 2002.
- 20.M. Lades, et al.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. on Computers, vol. 42, no. 3, pp. 300-311, March 1993.
- 21.J. Zhang, Y. Yan, and M. Lades: Face recognition: eigenface, elastic matching, and neural nets. Proceedings of the IEEE, vol. 85, no. 9, pp. 1423-1435, 1997.
- 22.L. Wiskott, J. Fellous, N. Kruger, and C. v. d. Malsburg: Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 775-779, July 1997.
- 23.L. Wiskott: Phantom faces for face analysis. Pattern Recognition, vol. 30, no. 6, pp. 837-846, 1997.
- 24.R. P. Wurtz: Object recognition robust under translations, deformations, and changes in background. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp. 769-775, July 1997.
- 25.B. Duc, S. Fischer, and J. Bigun: Face authentication with Gabor information on deformable graphs. IEEE Transactions on Image Processing, vol. 8, no. 4, pp. 504-516, Apr. 1999.
- 26.C. Kotropoulos, A. Tefas, and I. Pitas: Frontal face authentication using discriminating grids with morphological feature vectors. IEEE Transactions on Multimedia, vol. 2, no. 1, pp. 14-26, Mar. 2000.
- 27.C. Kotropoulos, A. Tefas, and I. Pitas: Morphological elastic graph matching applied to frontal face authentication under well-controlled and real conditions. Pattern Recognition, vol. 33, no. 12, pp. 31-43, Oct. 2000.
- 28.C. Kotropoulos, A. Tefas, and I. Pitas: Frontal face authentication using morphological elastic graph matching. IEEE Transactions on Image Processing, vol. 9, no. 4, pp. 555-560, Apr. 2000.
- 29.P. T. Jackway and M. Deriche: Scale-space properties of the multiscale morphological dilation-erosion. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18, no. 1, pp. 38-51, 1996.
- 30.A. Tefas, C. Kotropoulos, and I. Pitas: Face verification using elastic graph matching based on morphological signal decomposition. Signal Processing, vol. 82, no. 6, pp. 833-851, 2002.

31. N. Kruger: An algorithm for the learning of weights in discrimination functions using A priori constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 764–768, July 1997.
32. A. Tefas, C. Kotropoulos, and I. Pitas: Using support vector machines to enhance the performance of elastic graph matching for frontal face authentication. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, 2001.
33. Haitao Wang, Stan Z. Li, Yangsheng Wang, Weiwei Zhang: [Illumination Modeling and Normalization for Face Recognition](#). Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures. Nice, France. 2003.
34. Kyong I. Chang Kevin W. Bowyer Patrick J. Flynn: Face Recognition Using 2D and 3D Facial Data. In Proceedings Workshop in Multimodal User Authentication, pp. 25-32, Santa Barbara, California, December. 2003.
35. Kyong I. Chang Kevin W. Bowyer Patrick J. Flynn: An Evaluation of Multi-modal 2D+3D Face Biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.
36. Daniel González-Jiménez, Manuele Bicego, J. W. H. Tangelder, B. A. M Schouten, Onkar Ambekar, José Luis Alba-Castro, Enrico Grosso and Massimo Tistarelli: Distance Measures for Gabor Jets-Based Face Authentication: A Comparative Evaluation. In Proceedings 2nd Int.l Conference on Biometrics - ICB 2007, Seoul August 28-30 2007, pp 474-483, Springer LNCS 4642.
37. Bicego M., Brelstaff G., Brodo L., Grosso E., Lagorio A. and Tistarelli M.: Distinctiveness of faces: a computational approach. *ACM Transactions on Applied Perception*, Vol. 5, n. 2, 2008.
38. D.R. Kisku, A. Rattani, E. Grosso and M. Tistarelli: Face Identification by SIFT-based Complete Graph Topology. In Proceedings of IEEE Int.l Workshop on Automatic Identification Advanced Technologies (AutoId 2007) Alghero June 7-8 2007, pp 69-73.
39. A. Rattani, D.R. Kisku, M. Bicego, M. Tistarelli (2007) “Feature Level Fusion of Face and Fingerprint Biometrics”, Proc. of first IEEE Int.l Conference on Biometrics: Theory, Applications and Systems (BTAS 07), September 27th to 29th, 2007, Washington DC.
40. G. Gordon, M. Lewis: Face Recognition Using Video Clips and Mug Shots. In Proceedings of the Office of National Drug Control Policy (ONDCP) International Technical Symposium (Nashua, NH), October 1995.
41. Y. Li, S. Gong, and H. Liddell: Video-based online face recognition using identity surfaces. In Proceedings IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pages 40-46, Vancouver, Canada, July 2001.
42. Y. Li, S. Gong, and H. Liddell: Modelling faces dynamically across views and over time. In Proceedings IEEE International Conference on Computer Vision, pages 554-559, Vancouver, Canada, July 2001.
43. [Lucas, S.M.](#): Continuous n-tuple classifier and its application to real-time face recognition. In [IEEE Proceedings-Vision Image and Signal Processing](#) Vol 145, No. 5, October 1998, pp. 343.
44. [Lucas, S.M.](#), [Huang, T.K.](#): Sequence recognition with scanning N-tuple ensembles. In Proceedings [ICPR04](#) (III) pp 410-413.
45. [Eickeler, S.](#), [Müller, S.](#), [Rigoll, G.](#): Recognition of JPEG compressed face images based on statistical methods. [Image and Vision Computing, Vol 18](#), No. 4, March 2000, pp. 279-287.
46. [Raytchev, B.](#), [Murase, H.](#): Unsupervised recognition of multi-view face sequences based on pairwise clustering with attraction and repulsion. [Computer Vision and Image Understanding, Vol. 91](#), No. 1-2, July-August 2003, pp. 22-52.
47. [Raytchev, B.](#), [Murase, H.](#): VQ-Faces: Unsupervised Face Recognition from Image Sequences. In Proceedings [ICIP02](#) (II), pp 809-812.
48. [Raytchev, B.](#), [Murase, H.](#): Unsupervised Face Recognition from Image Sequences. In Proceedings [ICIP01](#)(I), pp 1042-1045.

49. [Zhou, S.](#), [Krueger, V.](#), [Chellappa, R.](#): Probabilistic recognition of human faces from video. [Computer Vision and Image Understanding, Vol. 91](#), No. 1-2, July-August 2003, pp. 214-245.
50. [Zhou, S.](#), [Krueger, V.](#), [Chellappa, R.](#): Face Recognition from Video: A Condensation Approach. In Proceedings [IEEE AFGRO2](#), pp 212-217.
51. [Zhou, S.](#), [Chellappa, R.](#): Probabilistic Human Recognition from Video. In Proceedings [ECCV02](#) (III), pp 681.
52. [Zhou, S.](#), [Chellappa, R.](#): A robust algorithm for probabilistic human recognition from video. In Proceedings [ICPR02](#) (I), pp 226-229.
53. [Zhou, S.](#), [Chellappa, R.](#): Rank constrained recognition under unknown illuminations. In Proceedings [AMFG03](#), pp 11-18.
54. [Zhou, S.K.](#), [Chellappa, R.](#), [Moghaddam, B.](#): Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters. [Image Processing, Vol 13](#), No. 11, November 2004, pp. 1491-1506.
55. [Zhou, S.K.](#), [Chellappa, R.](#), [Moghaddam, B.](#): Intra-personal kernel space for face recognition. In proceedings [IEEE AFGRO4](#), pp. 235-240.
56. [Zhou, S.K.](#), [Chellappa, R.](#): Multiple-exemplar discriminant analysis for face recognition. In Proceedings [ICPR04](#) (IV), pp 191-194.
57. [Zhou, S.K.](#), [Chellappa, R.](#): Probabilistic identity characterization for face recognition. In Proceedings [CVPR04](#) (II), pp 805-812.
58. A.J. Howell and H. Buxton: Towards Unconstrained Face Recognition from Image Sequences. In Proceeding. of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR'96), Killington, VT, pp.224-229, 1996.
59. Y. Li, S. Gong, H. Liddell: Support Vector Regression and Classification Based Multiview Face Detection and Recognition. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FGR'00), Grenoble, France, pp.300-305, 2000.
60. F. Roli and J. Kittler Eds. Multiple Classifier Systems. Springer Verlag, LNCS 2364, 2002.
61. B. Achermann and H. Bunke: Combination of Classifiers on the Decision Level for Face Recognition. Technical Report IAM-96-002, Institut für Informatik und angewandte Mathematik, Universität Bern, January 1996.
62. [Mou, D.](#), [Schweier, R.](#), [Rothermel, A.](#): Automatic Databases for Unsupervised Face Recognition. In Proceedings [FaceVideo04](#), pp 90.
63. [Song, X.](#), [Lin, C.Y.](#), [Sun, M.T.](#): Cross-Modality Automatic Face Model Training from Large Video Databases. In Proceedings [FaceVideo04](#), pp 91.
64. [Arandjelovic, O.](#), [Cipolla, R.](#): Face Recognition from Face Motion Manifolds using Robust Kernel Resistor-Average Distance. In Proceedings [FaceVideo04](#), pp 88.
65. [Aggarwal, G.](#), [Chowdhury, A.K.R.](#), [Chellappa, R.](#): A system identification approach for video-based face recognition. In Proceedings [ICPR04](#) (IV), pp 175-178.
66. [Matsui, A.](#), [Clippingdale, S.](#), [Uzawa, F.](#), [Matsumoto, T.](#): Bayesian face recognition using a Markov chain Monte Carlo method. In Proceedings [ICPR04](#) (III), pp 918-921.
67. [Clippingdale, S.](#), [Fujii, M.](#): Face Recognition for Video Indexing: Randomization of Face Templates Improves Robustness to Facial Expression. In Proceedings [VLBV03](#), pp 32-40.
68. [Clippingdale, S.](#), [Ito, T.](#): A Unified Approach to Video Face Detection, Tracking and Recognition. In Proceedings [ICIP99](#) (I), pp 662-666.
69. [Roark, D.A.](#), [O'Toole, A.J.](#), [Abdi, H.](#): Human recognition of familiar and unfamiliar people in naturalistic video. In Proceedings [AMFG03](#), pp 36-41.
70. [Gorodnichy, D.O.](#): Facial Recognition in Video. In Proceedings [AVBPA03](#), pp 505-514.
71. [Lee, K.C.](#), [Ho, J.](#), [Yang, M.H.](#), [Kriegman, D.J.](#): Video-based face recognition using probabilistic appearance manifolds. In Proceedings [CVPR03](#) (I), pp 313-320.

72. [Liu, X.](#), [Chen, T.](#): Video-based face recognition using adaptive hidden Markov models. In Proceedings [CVPR03](#) (I), pp 340-345.
73. [Huang, K.S.](#), [Trivedi, M.M.](#): Streaming face recognition using multicamera video arrays. In Proceedings [ICPR02](#) (IV), pp 213-216.
74. [Gross, R.](#), [Brajovic, V.](#): An Image Preprocessing Algorithm for Illumination Invariant Face Recognition. In Proceedings [AVBPA03](#), pp 10-18.
75. [Gross, R.](#), [Yang, J.](#), [Waibel, A.](#): Growing Gaussian Mixture Models for Pose Invariant Face Recognition. In Proceedings [ICPR00](#) (I), pp 1088-1091.
76. [Krüger, V.](#), [Gross, R.](#), [Baker, S.](#): Appearance-Based 3-D Face Recognition from Video. In Proceedings [DAGM02](#), pp 566.
77. [Krüger, V.](#), [Zhou, S.](#): Exemplar-Based Face Recognition from Video. In Proceedings, IEEE [AFGR02](#), pp 175-180.
78. [Shakhnarovich, G.](#), [Fisher, J.W.](#), [Darrell, T.J.](#): Face Recognition from Long-Term Observations. In Proceedings [ECCV02](#) (III), pp 851.
79. [Shakhnarovich, G.](#), [Viola, P.A.](#), [Moghaddam, B.](#): A unified learning framework for real time face detection and classification. In Proceedings IEEE [AFGR02](#), pp 14-21.
80. Li, Y.; Gong, S.; Liddell, H.: Video-based online face recognition using identity surfaces. In Proceedings IEEE ICCV Workshop on [RATFG01](#), pp 40-46.
81. [Weng, J.](#), [Evans, C.H.](#), [Hwang, W.S.](#): An Incremental Learning Method for Face Recognition under Continuous Video Stream. In Proceedings [IEEE AFGR00](#), pp 251-256.
82. [Ho, P.](#): Rotation Invariant Real-time Face Detection and Recognition System. [MIT-AI Memo 2001-010](#), May 31, 2001.
83. [Yamaguchi, O.](#), [Fukui, K.](#), [Maeda, K.](#): Face Recognition Using Temporal Image Sequence. In Proceedings [IEEE AFGR98](#), pp 318-323.
84. [Nagao, K.](#), [Sohma, M.](#): Weak Orthogonalization of Face and Perturbation for Recognition. In Proceedings [CVPR98](#), pp 845-852.
85. [Nagao, K.](#), [Sohma, M.](#): Recognizing faces by weakly orthogonalizing against perturbations. In Proceedings [ECCV98](#) (II), pp 613.
86. [Edwards, G.J.](#), [Taylor, C.J.](#), [Cootes, T.F.](#): Improving Identification Performance by Integrating Evidence from Sequences. In Proceedings [CVPR99](#) (I), pp 486-491.
87. [Cootes, T.F.](#), [Wheeler, G.V.](#), [Walker, K.](#), [Taylor, C.J.](#): Coupled-View Active Appearance Models. In Proceedings [BMVC00](#), pp 52-61.
88. [Edwards, G.J.](#), [Taylor, C.J.](#), and [Cootes, T.F.](#): Learning to Identify and Track Faces in Image Sequences. In Proceedings [IEEE AFGR98](#), pp 260-265.
89. Déniz, M. Castrillón, J. Lorenzo and M. Hernández: An Incremental Learning Algorithm for Face Recognition. In Proceedings Int.l Workshop on Biometric Authentication 2002, Copenhagen, Denmark, Springer Verlag, LNCS 2359, pp 1-9.
90. K. Fukui and O. Yamaguchi: Face recognition using multiviewpoint patterns for robot vision. In Proceedings International Symposium of Robotics Research, 2003.
91. Bicego, M., Grosso, E. and Tistarelli, M.: Person authentication from video of faces: a behavioral and physiological approach using Pseudo Hierarchical Hidden Markov Models. In Proceedings Intern.l Conference on Biometric Authentication 2006, Hong Kong, China, January 2006, pp 113-120, LNCS 3832.
92. Tistarelli, M., Bicego, M. and Grosso, E.: Dynamic face recognition: From human to machine vision. Image and Vision Computing: Special issue on Multimodal Biometrics, M. Tistarelli and J. Bigun ed.s, doi:10.1016/j.imavis.2007.05.006.
93. Phillips, J.J., Flynn, P., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Jaesik, M., Worek W.: Overview of the Face Recognition Grand Challenge. In Proceedings [CVPR05](#), pp 947-954, 2005.
94. Phillips, J.J., Flynn, P., Scruggs, T., Bowyer, K.W., Worek W.: Preliminary Face Recognition Grand Challenge Results. In Proceedings 7th International Conference on Automatic Face and Gesture Recognition, pp 15-24, 2006.

Discriminant Non-negative Matrix Factorization and Projected Gradients for Frontal Face Verification.

Irene Kotsia, Stefanos Zafeiriou, and Ioannis Pitas

Aristotle University of Thessaloniki, Department of Informatics, Box 451, 54124,
Greece

{ekotsia, dralbert, pitas}@aiaa.csd.auth.gr

<http://www.aiaa.csd.auth.gr>

Abstract. A novel *Discriminant Non-negative Matrix Factorization* (DNMF) method that uses projected gradients, is presented in this paper. The proposed algorithm guarantees the algorithm's convergence to a stationary point, contrary to the methods introduced so far, that only ensure the non-increasing behavior of the algorithm's cost function. The proposed algorithm employs some extra modifications that make the method more suitable for classification tasks. The usefulness of the proposed technique to the frontal face verification problem is also demonstrated.

Key words: Non-negative Matrix Factorization, projected gradients, frontal face verification.

1 Introduction

Over the past few years, the *Non-negative Matrix Factorization* (NMF) algorithm and its alternatives have been widely used, especially in facial image characterization and representation problems [3]. NMF aims at representing a facial image as a linear combination of basis images. Like *Principal Component Analysis* (PCA), NMF does not allow negative elements in either the basis images or the representation coefficients used in the linear combination of the basis images, thus representing the facial image only by additions of weighted basis images. The nonnegativity constraints introduced correspond better to the intuitive notion of combining facial parts to create a complete facial image.

In order to enhance the sparsity of NMF, many methods have been proposed for its further extension to supervised alternatives by incorporating discriminant constraints in the decomposition, the so-called DNMF or Fisher-NMF (FNMF) methods [3]. The intuitive motivation behind DNMF methods is to extract bases that correspond to discriminant facial regions and contain more discriminative information about them. A procedure similar to the one followed in the NMF decomposition [6] regarding the calculation of the update rules for the weights and the basis images was also used in the DNMF decomposition [3].

In this paper, a novel DNMF method is proposed that employs discriminant constraints on the classification features and not on the representation coefficients. Projected gradient methods are used for the optimization procedure to ensure that the limit point found will be a stationary point (similar methods have been applied to NMF [5]). Frontal face verification experiments were conducted and it has been demonstrated that the proposed method outperforms the other discriminant non-negative methods.

2 Discriminant Non-Negative Matrix Factorization Algorithms

2.1 Non-Negative Matrix Factorization

An image scanned row-wise is used to form a vector $\mathbf{x} = [x_1 \dots x_F]^T$ for the NMF algorithm. The basic idea behind NMF is to approximate the image \mathbf{x} by a linear combination of the basis images in $\mathbf{Z} \in \mathfrak{R}_+^{F \times M}$, whose coefficients are the elements of $\mathbf{h} \in \mathfrak{R}_+^M$ such that $\mathbf{x} \approx \mathbf{Z}\mathbf{h}$. Using the conventional least squares formulation, the approximation error $\mathbf{x} \approx \mathbf{Z}\mathbf{h}$ is measured in terms of $L(\mathbf{x}|\mathbf{Z}\mathbf{h}) \triangleq \|\mathbf{x} - \mathbf{Z}\mathbf{h}\|^2 = \sum_i (x_i - [\mathbf{Z}\mathbf{h}]_i)^2$. Another way to measure the error of the approximation is using the Kullback-Leibler (KL) divergence, $KL(\mathbf{x}|\mathbf{Z}\mathbf{h}) \triangleq \sum_i (x_i \ln \frac{x_i}{[\mathbf{Z}\mathbf{h}]_i} + [\mathbf{Z}\mathbf{h}]_i - x_i)$ [6] which is the most common error measure for all DNMF methods [3]. A limitation of KL-divergence is that it requires both \mathbf{x}_i and $[\mathbf{Z}\mathbf{h}]_i$ to be strictly positive (i.e., neither negative nor zero values are allowed).

In order to apply the NMF algorithm, the matrix $\mathbf{X} \in \mathfrak{R}_+^{F \times T} = [x_{ij}]$ should be constructed, where x_{ij} is the i -th element of the j -th image vector. In other words, the j -th column of \mathbf{X} is the facial image \mathbf{x}_j . NMF aims at finding two matrices $\mathbf{Z} \in \mathfrak{R}_+^{F \times M} = [z_{i,k}]$ and $\mathbf{H} \in \mathfrak{R}_+^{M \times T} = [h_{k,j}]$ such that:

$$\mathbf{X} \approx \mathbf{Z}\mathbf{H}. \quad (1)$$

After the NMF decomposition, the facial image \mathbf{x}_j can be written as $\mathbf{x}_j \approx \mathbf{Z}\mathbf{h}_j$, where \mathbf{h}_j is the j -th column of \mathbf{H} . Thus, the columns of the matrix \mathbf{Z} can be considered as basis images and the vector \mathbf{h}_j as the corresponding weight vector. The vector \mathbf{h}_i can be also considered as the projection of \mathbf{x}_j in a lower dimensional space.

The defined cost for the decomposition (1) is the sum of all KL divergences for all images in the database:

$$D(\mathbf{X}|\mathbf{Z}\mathbf{H}) = \sum_j KL(\mathbf{x}_j|\mathbf{Z}\mathbf{h}_j) = \sum_{i,j} \left(x_{i,j} \ln \left(\frac{x_{i,j}}{\sum_k z_{i,k} h_{k,j}} \right) + \sum_k z_{i,k} h_{k,j} - x_{i,j} \right). \quad (2)$$

The NMF factorization is the outcome of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{Z}, \mathbf{H}} D(\mathbf{X}|\mathbf{Z}\mathbf{H}) \text{ subject to} & \quad (3) \\ z_{i,k} \geq 0, h_{k,j} \geq 0, \sum_i z_{i,j} = 1, \forall j. & \end{aligned}$$

2.2 Discriminant Non-Negative Matrix Factorization

In order to formulate the DNMF algorithm, let the matrix \mathbf{X} that contains all the facial images be organized as follows. The j -th column of the database \mathbf{X} is the ρ -th image of the r -th image class. Thus, $j = \sum_{i=1}^{r-1} N_i + \rho$, where N_i is the cardinality of the image class i . The r -th image class could consist of one person's facial images, for face recognition and verification problems. The vector \mathbf{h}_j that corresponds to the j -th column of the matrix \mathbf{H} , is the coefficient vector for the ρ -th facial image of the r -th class and will be denoted as $\boldsymbol{\eta}_\rho^{(r)} = [\eta_{\rho,1}^{(r)} \dots \eta_{\rho,M}^{(r)}]^T$. The mean vector of the vectors $\boldsymbol{\eta}_\rho^{(r)}$ for the class r is denoted as $\boldsymbol{\mu}^{(r)} = [\mu_1^{(r)} \dots \mu_M^{(r)}]^T$ and the mean of all classes as $\boldsymbol{\mu} = [\mu_1 \dots \mu_M]^T$. Then, the within-class scatter matrix for the coefficient vectors \mathbf{h}_j is defined as:

$$\mathbf{S}_w = \sum_{r=1}^K \sum_{\rho=1}^{N_r} (\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})(\boldsymbol{\eta}_\rho^{(r)} - \boldsymbol{\mu}^{(r)})^T \quad (4)$$

whereas the between-class scatter matrix is defined as:

$$\mathbf{S}_b = \sum_{r=1}^K N_r (\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu})^T. \quad (5)$$

The matrix \mathbf{S}_w defines the scatter of the sample vector coefficients around their class mean. The dispersion of samples that belong to the same class around their corresponding mean should be as small as possible. A convenient measure for the dispersion of the samples is the trace of \mathbf{S}_w . The matrix \mathbf{S}_b denotes the between-class scatter matrix and defines the scatter of the mean vectors of all classes around the global mean $\boldsymbol{\mu}$. Each class must be as far as possible from the other classes. Therefore, the trace of \mathbf{S}_b should be as large as possible.

To formulate the DNMF method [3], discriminant constraints have been incorporated in the NMF decomposition inspired by the minimization of the Fisher's criterion [3]. The DNMF cost function is given by:

$$D_d(\mathbf{X}||\mathbf{ZH}) = D(\mathbf{X}||\mathbf{ZH}) + \gamma \text{tr}[\mathbf{S}_w] - \delta \text{tr}[\mathbf{S}_b] \quad (6)$$

where γ and δ are non-negative constants. The update rules that guarantee a non-increasing behavior of (6) for the weights $h_{k,j}$ and the bases $z_{i,k}$, under the constraints of (2), can be found in [3]. Unfortunately, the update rules only guarantee a non-increasing behavior for (6) and do not ensure that the limit point will be stationary.

3 Projected Gradient Methods for Discriminant Non-Negative Matrix Factorization

Let $\mathbf{E} = \mathbf{X} - \mathbf{ZH}$ be the error signal of the decomposition. The modified optimization problem should minimize:

$$D_p(\mathbf{X}||\mathbf{ZH}) = \|\mathbf{E}\|_F^2 + \gamma \text{tr}[\tilde{\mathbf{S}}_w] - \delta \text{tr}[\tilde{\mathbf{S}}_b], \quad (7)$$

under non-negativity constraints, where $\|\cdot\|_F$ is the Frobenius norm. The within-class scatter matrix $\tilde{\mathbf{S}}_w$ and the between-scatter scatter matrix $\tilde{\mathbf{S}}_b$ are defined using the vectors $\tilde{\mathbf{x}}_j = \mathbf{Z}^T \mathbf{x}_j$ and the definitions of the scatter matrices in (4) and (5).

The minimization of (7) subject to nonnegative constraints yields the new discriminant nonnegative decomposition. The new optimization problem is the minimization of (7) subject to non-negative constraints for the weights matrix \mathbf{H} and the bases matrix \mathbf{Z} . This optimization problem will be solved using projected gradients in order to guarantee that the limit point will be stationary. In order to find the limit point, two functions are defined:

$$f_{\mathbf{Z}}(\mathbf{H}) = D_p(\mathbf{X}|\|\mathbf{ZH}) \text{ and } f_{\mathbf{H}}(\mathbf{Z}) = D_p(\mathbf{X}|\|\mathbf{ZH}) \quad (8)$$

by keeping \mathbf{Z} and \mathbf{H} fixed, respectively.

The projected gradient method used in this paper, successively optimizes two subproblems [5]:

$$\min_{\mathbf{Z}} f_{\mathbf{H}}(\mathbf{Z}) \text{ subject to, } z_{i,k} \geq 0, \quad (9)$$

and

$$\min_{\mathbf{H}} f_{\mathbf{Z}}(\mathbf{H}) \text{ subject to, } h_{k,j} \geq 0. \quad (10)$$

The method requires the calculation of the first and the second order gradients of the two functions in (8):

$$\begin{aligned} \nabla f_{\mathbf{Z}}(\mathbf{H}) &= \mathbf{Z}^T (\mathbf{ZH} - \mathbf{X}) \\ \nabla^2 f_{\mathbf{Z}}(\mathbf{H}) &= \mathbf{Z}^T \mathbf{Z} \\ \nabla f_{\mathbf{H}}(\mathbf{Z}) &= (\mathbf{ZH} - \mathbf{X}) \mathbf{H}^T + \gamma \nabla \text{tr}[\tilde{\mathbf{S}}_w] - \delta \nabla \text{tr}[\tilde{\mathbf{S}}_b] \\ \nabla^2 f_{\mathbf{H}}(\mathbf{Z}) &= \mathbf{HH}^T + \gamma \nabla^2 \text{tr}[\tilde{\mathbf{S}}_w] - \delta \nabla^2 \text{tr}[\tilde{\mathbf{S}}_b]. \end{aligned} \quad (11)$$

The projected gradient DNMF method is an iterative method that is comprised of two main phases. These two phases are iteratively repeated until the ending condition is met or the number of iterations exceeds a given number. In the first phase, an iterative procedure is followed for the optimization of (9), while in the second phase, a similar procedure is followed for the optimization of (10). In the beginning, the bases matrix $\mathbf{Z}^{(1)}$ and the weight matrix $\mathbf{H}^{(1)}$ are initialized either randomly or by using structured initialization [7], in such a way that their entries are nonnegative. The regularization parameters γ and δ that are used to balance the trade-off between accuracy of the approximation and discriminant decomposition of the computed solution and their selection is typically problem dependent.

3.1 Solving the Subproblem (9)

Consider the subproblem of optimizing with respect to \mathbf{Z} , while keeping the matrix \mathbf{H} constant. The optimization is an iterative procedure that is repeated until $\mathbf{Z}^{(t)}$ becomes a stationary point of (9). In every iteration, a proper step size a_t is required to update the matrix $\mathbf{Z}^{(t)}$. When a proper update is found, the stationarity condition is checked and, if met, the procedure stops.

Update the matrix \mathbf{Z} For a number of iterations $t = 1, 2, \dots$ the following updates are performed [5]:

$$\mathbf{Z}^{(t+1)} = P \left[\mathbf{Z}^{(t)} - a_t \nabla f_{\mathbf{H}}(\mathbf{Z}^{(t)}) \right] \quad (12)$$

where $a_t = \beta^{g_t}$ and g_t is the first non-negative integer such that:

$$f_{\mathbf{H}}(\mathbf{Z}^{(t+1)}) - f_{\mathbf{H}}(\mathbf{Z}^{(t)}) \leq \sigma \left\langle \nabla f_{\mathbf{H}}(\mathbf{Z}^{(t)}), \mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)} \right\rangle. \quad (13)$$

The projection rule $P[\cdot] = \max[\cdot, 0]$ refers to the elements of the matrix and guarantees that the update will not contain any negative entries. The operator $\langle \cdot, \cdot \rangle$ is the inner product between matrices defined as:

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_i \sum_j a_{i,j} b_{i,j} \quad (14)$$

where $[\mathbf{A}]_{i,j} = a_{i,j}$ and $[\mathbf{B}]_{i,j} = b_{i,j}$. The condition (13) ensures the sufficient decrease of the $f_{\mathbf{H}}(\mathbf{Z})$ function values per iteration. Since the function $f_{\mathbf{H}}$ is quadratic in terms of \mathbf{Z} , the inequality (13) can be reformulated as:

$$(1 - \sigma) \left\langle \nabla f_{\mathbf{H}}(\mathbf{Z}^{(t)}), \mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)} \right\rangle + \frac{1}{2} \left\langle \mathbf{Z}^{(t+1)} - \mathbf{Z}^{(t)}, \nabla^2 f_{\mathbf{H}}(\mathbf{Z}^{(t+1)}) \right\rangle \leq 0 \quad (15)$$

which is the actual condition checked.

The search of a proper value for a_t is the most time consuming procedure, thus, as few iteration steps as possible are desired. Several procedures have been proposed for the selection and update of the a_t values [8]. The Algorithm 4 in [5] has been used in our experiments and β , σ are chosen to be equal to 0.1 and 0.01 ($0 < \beta < 1$, $0 < \sigma < 1$), respectively. The choice of σ has been thoroughly studied in [5, 8]. During experiments it was observed that a smaller value of β reduces more aggressively the step size, but it may also result in a step size that is too small. The search for a_t is repeated until the point $\mathbf{Z}^{(t)}$ becomes a stationary point.

Check of Stationarity In this step it is checked whether or not in the limit point the first order derivatives are close to zero (stationarity condition). A commonly used condition to check the stationarity of a point is the following [8]:

$$\|\nabla^P f_{\mathbf{H}}(\mathbf{Z}^{(t)})\|_F \leq \epsilon_{\mathbf{Z}} \|\nabla f_{\mathbf{H}}(\mathbf{Z}^{(1)})\|_F \quad (16)$$

where $\nabla^P f_{\mathbf{H}}(\mathbf{Z})$ is the projected gradient for the constraint optimization problem defined as:

$$[\nabla^P f_{\mathbf{H}}(\mathbf{Z})]_{i,k} = \begin{cases} [\nabla f_{\mathbf{H}}(\mathbf{Z})]_{i,k} & \text{if } z_{i,k} > 0 \\ \min(0, [\nabla f_{\mathbf{H}}(\mathbf{Z})]_{i,k}) & z_{i,k} = 0. \end{cases} \quad (17)$$

and $0 < \epsilon_{\mathbf{Z}} < 1$ is the predefined stopping tolerance. A very low $\epsilon_{\mathbf{Z}}$ (i.e., $\epsilon_{\mathbf{Z}} \approx 0$) leads to a termination after a large number of iterations. On the other hand, a tolerance close to one will result in a premature iteration termination.

3.2 Solving the Subproblem (10)

A similar procedure should be followed in order to find a stationary point for the subproblem (10) while keeping fixed the matrix \mathbf{Z} and optimizing in respect of \mathbf{H} . A value for a_t is iteratively sought and the weight matrix is updated according to:

$$\mathbf{H}^{(t+1)} = P \left[\mathbf{H}^{(t)} - a_t \nabla f_{\mathbf{Z}}(\mathbf{H}^{(t)}) \right] \quad (18)$$

until the function $f_{\mathbf{Z}}(\mathbf{H})$ value is sufficient decreased and the following inequality holds $\langle a, b \rangle$:

$$(1 - \sigma) \left\langle \nabla f_{\mathbf{Z}}(\mathbf{H}^{(t)}), \mathbf{H}^{(t+1)} - \mathbf{H}^{(t)} \right\rangle + \frac{1}{2} \left\langle \mathbf{H}^{(t+1)} - \mathbf{H}^{(t)}, \nabla^2 f_{\mathbf{Z}}(\mathbf{H}^{(t+1)}) \right\rangle \leq 0. \quad (19)$$

This procedure is repeated until the limit point $\mathbf{H}^{(t)}$ is stationary. The stationarity is checked using a similar criterion to (16), i.e.:

$$\|\nabla^P f_{\mathbf{Z}}(\mathbf{H}^{(t)})\|_F \leq \epsilon_{\mathbf{H}} \|\nabla f_{\mathbf{Z}}(\mathbf{H}^{(1)})\|_F \quad (20)$$

where $\epsilon_{\mathbf{H}}$ is the predefined stopping tolerance for this subproblem.

3.3 Convergence Rule

The procedure followed for the minimization of the two subproblems, in Sections 3.1 and 3.2, is iteratively followed until the global convergence rule is met:

$$\|\nabla f(\mathbf{H}^{(t)})\|_F + \|\nabla f(\mathbf{Z}^{(t)})\|_F \leq \epsilon \left(\|\nabla f(\mathbf{H}^{(1)})\|_F + \|\nabla f(\mathbf{Z}^{(1)})\|_F \right) \quad (21)$$

which checks the stationarity of the solution pair $\mathbf{H}^{(t)}, \mathbf{Z}^{(t)}$.

4 Experimental Results

The proposed DNMF method will be denoted as Projected Gradient DNMF (PGDNMF) from now onwards. The experiments were conducted in the XM2VTS database using the protocol described in [12]. The images were aligned semi-automatically according to the eyes position of each facial image using the eye coordinates. The facial images were down-scaled to a resolution of 64×64 pixels. Histogram equalization was used for the normalization of the facial image luminance.

The XM2VTS database contains 295 subjects, 4 recording sessions and two shots (repetitions) per recording session. It provides two experimental setups namely, Configuration I and Configuration II [12]. Each configuration is divided into three different sets: the training set, the evaluation set and the test set. The training set is used to create client and impostor models for each person. The evaluation set is used to learn the verification decision thresholds. In case of multimodal systems, the evaluation set is also used to train the fusion manager

[12]. For both configurations the training set has 200 clients, 25 evaluation impostors and 70 test impostors. The two configurations differ in the distribution of client training and client evaluation data. For additional details concerning the XM2VTS database an interested reader can refer to [12].

The experimental procedure followed was the one also used in [3]. For comparison reasons the same methodology using the Configuration I of the XM2VTS database was used. The performance of the algorithms is quoted by the Equal Error Rate (EER) which is the scalar figure of merit that is often used to judge the performance of a verification algorithm. An interested reader may refer to [12, 3] for more details concerning the XM2VTS protocol and the experimental procedure followed. In Figure 1, the verification results are shown for the various tested approaches, NMF [6], LNMF [11], DNMF [3], Class Specific DNMF [3], PCA [9], PCA plus LDA [10] and the proposed PGDNMF. EER is plotted versus the dimensionality of the new lower dimension space. As can be seen, the proposed PGDNMF algorithm outperforms (giving a best $EER \approx 2.0\%$) all the other part-based approaches and PCA. The best performance of LDA has been 1.7% which very close to the best performance of PGDNMF.

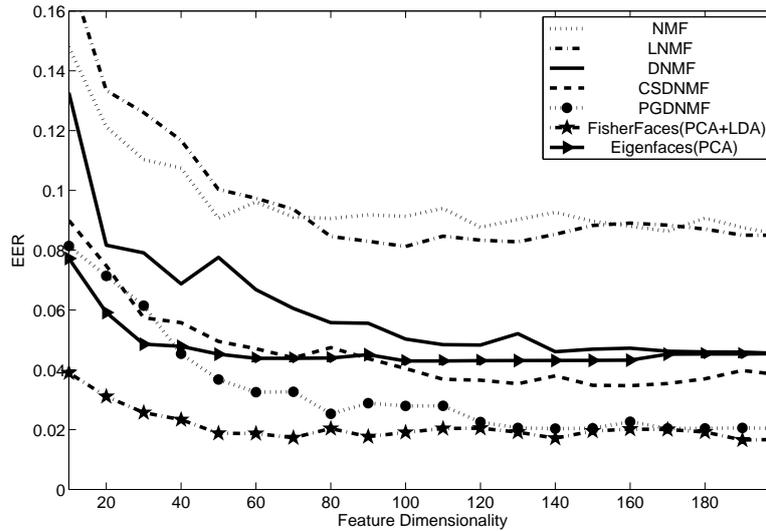


Fig. 1. EER for Configuration I of XM2VTS versus dimensionality.

5 Conclusions

A novel DNMF method has been proposed based on projected gradients. The incorporated discriminant constraints focus on the actual features used for classification and not on the weight vectors of the decomposition. Moreover, we have applied projected gradients in order to assure that the limit point is stationary. The proposed technique has been applied in supervised facial feature extraction for face verification, where it was shown that it outperforms several others subspace methods.

Acknowledgments. This work has been partially supported by the COST 2101 "Biometrics for Identity Documents and Smart Cards", www.cost2101.org.

References

1. D.D. Lee and H.S. Seung : Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791 (1999)
2. I. Buciu and I. Pitas : Application of non-negative and local non negative matrix factorization to facial expression recognition. In: ICPR 2004, pp. 288–291. Cambridge, United Kingdom (2004)
3. S. Zafeiriou, A. Tefas, I. Buciu and I. Pitas: Exploiting Discriminant Information in Nonnegative Matrix Factorization With Application to Frontal Face Verification. *IEEE Transactions on Neural Networks*, vol. 17, num. 3, pp. 683–695 (2006)
4. M. Kirby and L. Sirovich : Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces. *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 12, num. 1, 103–108 (1990)
5. C.-J. Lin: Projected gradient methods for non-negative matrix factorization. Technical report, Department of Computer Science, National Taiwan University (2005)
6. D.D. Lee and H.S. Seung : Algorithms for Non-negative Matrix Factorization. In: NIPS 2000, pp. 556–562.
7. I. Buciu, N. Nikolaidis and I. Pitas : On the initialization of the DNMF algorithm. In: Proc. of 2006 IEEE International Symposium on Circuits and Systems, Kos, Greece (2006).
8. C.-J. Lin and J.J. More: Newton’s method for large-scale bound constrained problems. *SIAM Journal on Optimization*, vol. 9, pp. 1100–1127 (1999)
9. M. Turk and A. P. Pentland: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86 (1991)
10. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman : Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, num. 7, pp. 711–720 (1997)
11. S.Z. Li, X.W. Hou and H.J. Zhang: Learning Spatially Localized, Parts-Based Representation. In: CVPR 2001, Kauai, HI, USA (2001).
12. K. Messer, J. Matas, J.V. Kittler, J. Luettin and G. Maitre : XM2VTSDB: The Extended M2VTS Database. In: AVBPA’99, pp. 72–77, Washington, DC, USA (1999).

On the Discrimination Capabilities of Speech Cepstral Features

A. Malegaonkar¹, A. Ariyaeinia, P. Sivakumaran, S. Pillay

University of Hertfordshire, College Lane, Hatfield, Hertfordshire, AL10 9AB, UK
amalegaonkar@trinityconvergence.com, {a.m.ariyaeinia, p.sivakumaran,
s.g.pillay}@herts.ac.uk

Abstract. In this work, the discrimination capabilities of speech cepstra for text and speaker related information are investigated. For this purpose, Bhattacharya distance metric is used as the measure of discrimination. The scope of the study covers static and dynamic cepstra derived using the linear prediction analysis (LPCC) as well as mel-frequency analysis (MFCC). The investigations also include the assessment of the linear prediction-based mel-frequency cepstral coefficients (LP-MFCC) as an alternative speech feature type. It is shown experimentally that whilst contaminations in speech unfavourably affect the performance of all types of cepstra, the effects are more severe in the case of MFCC. Furthermore, it is shown that with a combination of static and dynamic features, LP-based mel-frequency cepstra (LP-MFCC) exhibit the best discrimination capabilities in almost all experimental cases.

1 Introduction

Cepstra are the most commonly used features in speech related recognition tasks [1-4]. By definition, cepstrum of a given signal is obtained using homomorphic filtering which converts convolved source and filter impulse responses to linear summations [5]. An approach to extracting cepstral features from speech is that of first computing the speech linear prediction coefficients and then converting these to cepstral coefficients. Feature parameters obtained in this way are called linear prediction-based cepstral coefficients (LPCC) [5]. A second widely used method involves applying a mel-scale filter-bank function to the speech spectrum. The resultant feature parameters are referred to as mel-scale cepstral coefficients (MFCC) [3]. There are other types of cepstra that can be obtained through some variations of, or additional processing in, the above approaches. Examples of these are perceptual linear prediction coefficients (PLP) and linear filter bank cepstral coefficients (LFCC) [5]. Since LPCC and MFCC are the most widely used speech features, it is natural to focus the work on these. The indications from our initial study have been that each of these two feature types may possess certain superior discriminative characteristics, depending on the experimental conditions and the attribute considered. Therefore, in

¹ During the course of this work, Malegaonkar was with the University of Hertfordshire

an attempt to capture the benefits of each of these two commonly used classes of cepstra in one parametric representation, linear prediction-based, mel-frequency cepstral coefficients (LP-MFCC) are also considered in this study as an alternative feature type. The approach to extracting this class of speech features is given later in this paper. It should be pointed out that, whilst the idea behind LP-MFCC has been presented in some other studies [6-7], there is very limited information in the literature about the discrimination capabilities of this feature type [6].

The previous studies on the usefulness of various types of cepstra have been confined to individual applications. Examples are speaker recognition [1-2], speech recognition [3] and emotion recognition [4]. There have also been investigations into the usefulness of combining other features with cepstra for improving the performance, but again in a particular application only [8]. An important feature lacking in these studies is that of identifying the influence of the underlying experimental conditions on the outcomes. For instance, it is not known how variation in gender can affect the relative performance of different types of cepstra in text-dependent speaker recognition. Additionally, studies carried out to date have not been based on the same experimental setup or conditions. As a result, to date, the literature lacks information on the relative discrimination capabilities of different types of cepstra, in terms of individual classes of information contained in speech.

In general, the discrimination of any two sets of cepstral data can be achieved by assessing the divergence or distance between their distributions. Assuming that the distribution of such data is Gaussian, there are various metrics that can be used for this purpose. Although the underlying distribution of multidimensional cepstral data deviates from the Gaussian assumption, many speech applications such as speaker tracking and speaker segmentation use the Gaussian assumption of distribution for speech cepstral features. This assumption is reasonable as speech cepstra have unimodal distributions resembling Gaussians [5]. Additionally, when comparing the distributions of two sets of cepstral data directly using Gaussian-based measures, the exact Gaussian assumption of the distributions will not have a significant effect on the outcome as this is applied to both datasets. Examples of Gaussian-based comparative measures are Euclidean distance, Mahalanobis distance, and various other statistical measures [9]. Some of these measures show insensitivity towards particular data statistics, while some fail under certain conditions. For example, if the Euclidean distance between the means of two Gaussian distributions is used as a distance measure, then the covariance information is totally ignored. On the other hand, the F-Ratio which is a useful metric in terms of variance information has the drawback of being insensitive to mean statistics of data [5, 10, 11]. Amongst various measures, it is reported that Bhattacharya distance metric is well suited to the classification purpose [12] which is the main task in this study. As indicated in the study of speaker tracking in [9], for certain other purposes, it may be that the use of a different type of metric is advantageous. However, the nature of task in the present study together with the characteristics of Bhattacharya measure provides a strong justification for the deployment of this metric. The suitability of Bhattacharya distance is further discussed in Section 2.

The rest of this paper is organised as follows. Section 2 gives an overview of Bhattacharya distance as a discriminative measure. Section 3 details the experimental data and procedures together with various configurations used for the purpose of

investigations. The experimental results together with the discussions of these are presented in Section 4, and overall conclusions are given in Section 5.

2 Bhattacharya Distance

Bhattacharya distance for normal distributions is a very convenient measure for evaluating the class-separation capability [12]. If the multivariate data from two classes of A and B are normally distributed with statistics $A \in N(\mu_A, \Sigma_A)$ and $B \in N(\mu_B, \Sigma_B)$, where $N(\mu_i, \Sigma_i)$ are mean and covariance parameters for data I , then the metric is given as [12]

$$M = \frac{1}{8} \left[(\mu_A - \mu_B) \left(\frac{\Sigma_A + \Sigma_B}{2} \right)^{-1} (\mu_A - \mu_B)^T \right] + \frac{1}{2} \ln \left[\frac{\left| \frac{(\Sigma_A + \Sigma_B)}{2} \right|}{|\Sigma_A|^{1/2} |\Sigma_B|^{1/2}} \right], \quad (1)$$

where T is the transpose operation. The first term in this metric measures the distance between μ_A and μ_B normalised by the average covariance matrix, whilst the second term measures the distance due to covariance differences in data classes A and B. Hence, the first term gives class separation capability due to mean statistics from the two data sets and second term gives the separation capability due to covariance structures. Metric M itself gives the overall class separation capability.

The divergence between Gaussian distributions of the two data sets can act as another suitable distance measure [12]. However, a main drawback of using the divergence measure is due to its weak association with the Bayes error for the classification purposes [12]. The formulation of the divergence measure is based on various approximations in obtaining discrimination criterion for the two class problem. Hence in this work, Bhattacharya distance is adopted.

3 Experimental Procedures

3.1 Speech Data

For the purpose of experiments, TIMIT database is adopted. The advantage of using this database is that it is phonetically rich, and is recorded under clean background conditions. This reduces variability due to background environments and ensures that Gaussian statistics of cepstra are not pre-contaminated by noise. This database is also useful for studying the effects of the addition of noise to speech in a controlled manner. The database consists of speech material from 192 females and 438 males,

each with 10 utterances. In this work, material from 192 males and 192 females is used with 'sa1' and 'sx1' utterances.

3.2 Feature Parameter Representation

The extraction of cepstral parameters in this study is based on first pre-emphasising the input speech data using a first-order digital filter and then segmenting it into 20 ms frames at intervals of 10 ms using a Hamming window. A voice activity detection algorithm is then used to discard frames containing silence only. For each frame, 16 LPCC are obtained via a linear prediction analysis. To extract MFCC, the speech spectrum for each frame is weighted by a mel-scale filter bank. This filter bank consists of 26 triangular filters for the considered sampling frequency of 16 kHz. The discrete cosine transformation of the log magnitude outputs of these filters gives 16 MFCC for that speech frame. The extraction of LP-MFCC is based on first computing 16 LP coefficients for each frame. The above-stated perceptual processing is then deployed to obtain 16 mel-frequency coefficients from the LP spectrum. For each type of cepstra, a polynomial fit method is used to obtain 16 delta coefficients [5].

3.3 Experimental Configurations

Tests are carried out separately using various configurations as follows.

1. In this configuration LPCC, MFCC and LP-MFCC are assessed for their text-based discrimination capabilities under clean conditions. For each speaker, the Bhattacharya distance between Gaussian distributions of cepstra obtained using 'sa1' and 'sx1' utterances is computed. The tests are carried out separately for each gender, using static coefficients with and without delta coefficients. The mean of the Bhattacharya distance is estimated with 95 % confidence interval in each case.
2. Here, LPCC, MFCC and LP-MFCC are assessed for their speaker separation capabilities under clean conditions. The Bhattacharya distance is applied to the Gaussian distributions of cepstra obtained from pairs of speakers speaking the 'sa1' utterance. The tests are carried out separately within each gender group as well as across the genders, using static coefficients with and without delta coefficients. The mean of Bhattacharya distance in each case is estimated with 95 % confidence interval.
3. In this configuration, the tests are the same as in 1, but here the speech data is contaminated with Gaussian white noise. To examine the effects of contamination level, a range of signal-to-noise ratios (SNR) are used. These are 20 dB, 15 dB and 10 dB.
4. The tests are the same as in 2 but the speech data is contaminated with various levels of Gaussian white noise, producing different signal to noise ratios as detailed in 3.

4 Results and Discussions

The required keys for the interpretation of all the results presented below are as follows.

“M:” - within male speakers, “F:” - within female speakers, “M/F:” - between Male and female speakers. Vertical lines at the top of each bar represent the 95 % confidence interval values.

4.1 Text Separation Capabilities of Cepstra for Clean Speech

According to the results in Figure 1, the static MFCC and LPCC have almost the same capabilities for the separation of textual information. The results also show the advantages offered by using delta coefficients. Figure 1 further indicates that the performance of MFCC+delta is closely followed by that of LPCC+delta. This result is consistent with the pervious results obtained in speech recognition experiments [3]. It is interesting to note that LP-MFCC are noticeably better than LPCC and MFCC. This difference in performance appears to become significant when delta parameters are appended to the static features. Additionally, it is seen that, with or without using delta coefficients, the textual separation capabilities in the case of female speakers are always below those for male speakers.

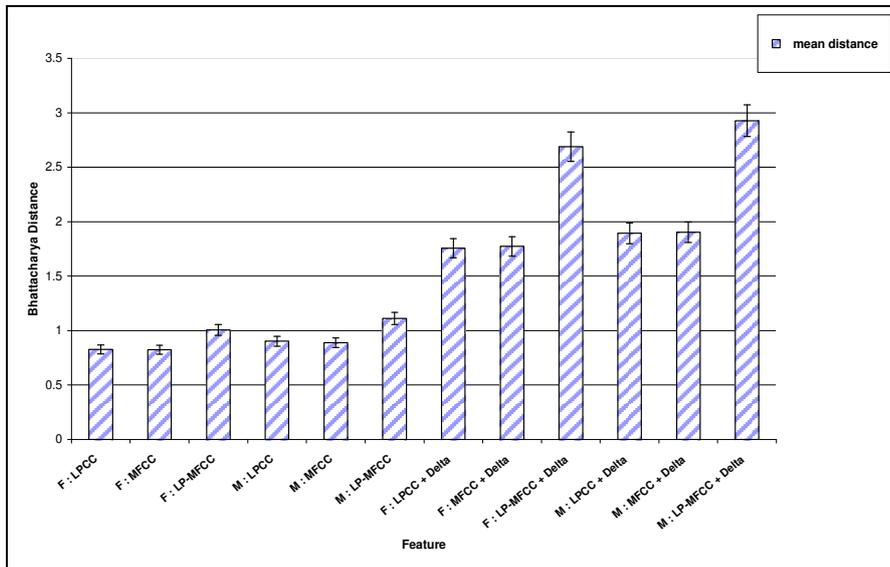


Fig. 1. Text separation capabilities of cepstra: experimental results based on configuration 1.

4.2 Speaker Separation Capabilities of Cepstra for Clean Speech

It can be seen that, with clean speech and the same gender speakers, LPCC offer only slightly better speaker separation capabilities than MFCC. However, a more noticeable difference in performance in favour of LPCC is observed for the combination of static and dynamic features. In the case of cross-gender tests, however, MFCC exhibit better discrimination capabilities than LPCC. This appears to be the case for both static features, as well as combined static-dynamic features. In terms of static features only, the capabilities offered by LP-MFCC appear to be between those of LPCC and MFCC, for both within gender and cross-gender tests. However, it should be noted that the performance of all three feature types improves considerably by appending delta parameters to static coefficients. In this case, the best performance is offered by LP-MFCC. Another interesting aspect of the results in Figure 2 is that, for every feature type, the discrimination achievable for male speakers is better than that for female speakers.

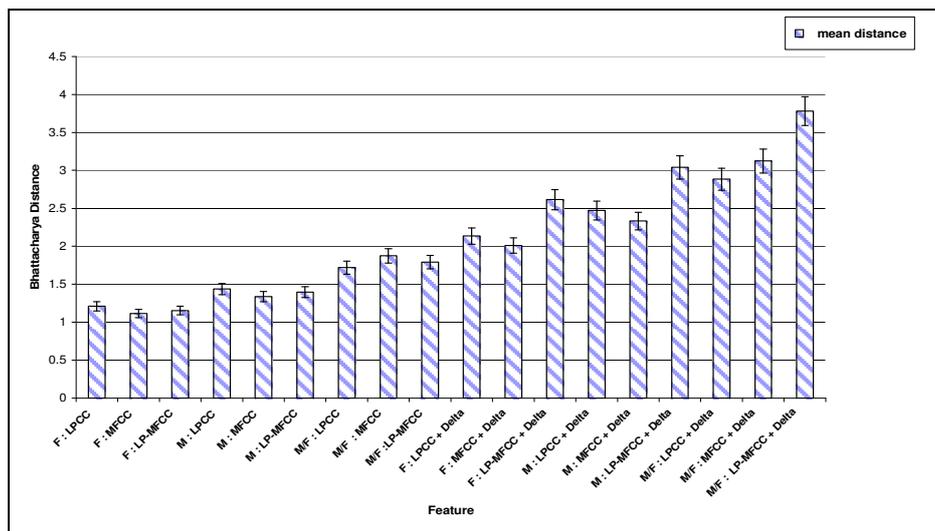


Fig. 2. Speaker separation capabilities of cepstra for clean speech: results of the experiments based on configuration 2.

4.3 Text Separation Capabilities of Cepstra for Noisy Speech

As seen in Figure 3, the textual separation capabilities for both genders deteriorate with decreasing SNR. It is also noted that the adverse effects of the additive noise are more considerable in the case of MFCC features. This imbalance in effects appears to even reverse the relative performance of LPCC and MFCC in favour of the former when a combination of static and dynamic features is used. As a result, for both

genders and all levels of contamination, LPCC features exhibit better discrimination capabilities than MFCC. As observed in Figure 3, although the LP-MFCC performance is also affected by noise, the effectiveness of this feature type is consistently better than that of LPCC and MFCC.

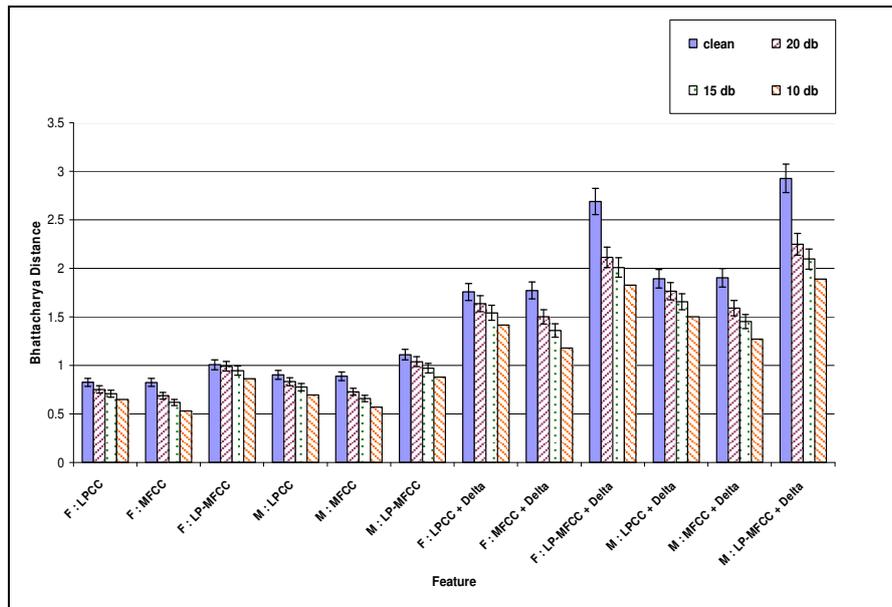


Fig. 3. Text separation capabilities of cepstra under Noisy Conditions: results of the experiments based on configuration 3.

4.4 Speaker Separation Capabilities of Cepstra for Noisy Speech

The results in Figure 4 show the effects of additive noise on speaker separation capabilities of LPCC, MFCC and LP-MFCC. It can be observed that these results are consistent with those in Figure 3. That is, as discussed above, the additive noise has more noticeable adverse effects on the results for MFCC features to the extent that better performance is obtained with LPCC in cross-gender tests. This relative performance, as observed, is regardless of using static features or combined static-dynamic features. It is also noted that, the decrease in SNR reduces the performance of the two categories of features at different rates. Therefore, the gap in the relative performance continuously increases (with noise level) in favour of LPC features. The results also show that LP-MFCC features exhibit a better level of robustness against noise than MFCC. It is observed that in cross-gender tests in the presence of noise, LP-MFCC perform better than the other two feature types. In single-gender experiments based on static features only, comparable performance is observed for LPCC and LP-MFCC.

In experiments based on static+delta features, LP-MFCC continues to offer better overall effectiveness for almost all noise levels. The only exception is the experiments with male speakers where comparable performance is obtained with LPCC.

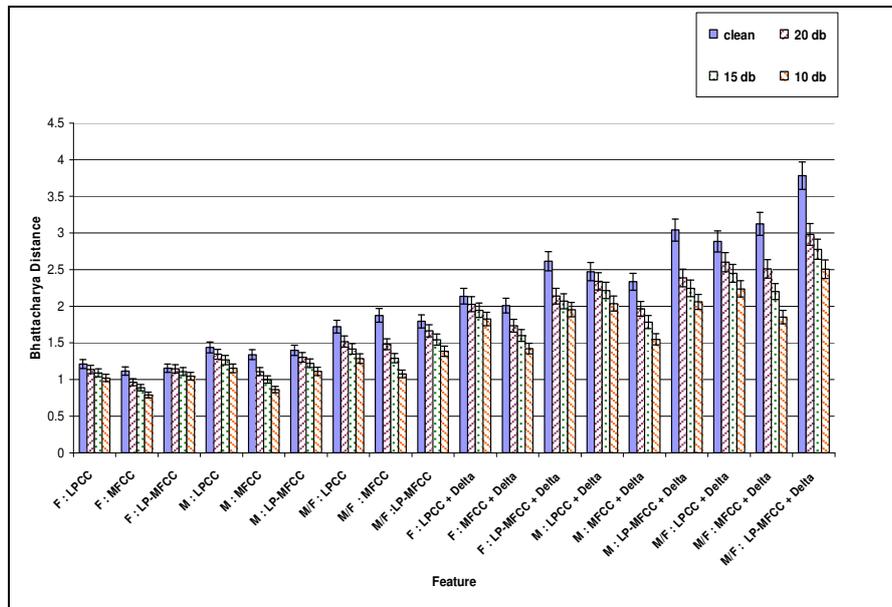


Fig. 4. Speaker separation capabilities under Noisy Conditions: results of the experiments based on configuration 4.

5 Conclusions

The discrimination capabilities of cepstra in terms of text and speaker identity have been investigated. By making the study independent of any particular application, attempts have been made to avoid the influence of application-specific conditions and parameters on the outcomes. For the purpose of this study, two commonly used types of cepstra (LPCC and MFCC) together with LP-based mel-frequency cepstra (LP-MFCC) have been investigated. The evaluations of discrimination capabilities have been conducted using the Bhattacharya distance. Based on the experimental results, it is concluded that the information discrimination capabilities in all categories of cepstra show dependence on the speaker gender and also on the test conditions. In terms of speaker discrimination, LPCC appear to exhibit marginally better performance when the speakers are of the same gender. In the case of cross-gender speaker discrimination, the experimental results have revealed that the MFCC features provide better performance than the LPCC features. The experimental results have also shown that, as expected, the speech contamination due to white noise affects the

performance of all types of cepstra. It is, however, observed that the effect is more significant in the case of MFCC.

The experiments conducted suggest that some useful discriminative characteristics of LPCC and MFCC are captured in LP-MFCC. This is evident by the fact that, in every case, LP-MFCC are found to either offer the best performance or to be almost as effective as the best performer.

In general, the use of delta coefficients in addition to static parameters has been found to considerably improve the separation capabilities of cepstra. In this case, the use of LP-MFCC appears to be advantageous as it provides the best performance in almost all cases. Although the study has been confined to cepstra, the approach adopted can also be used for assessing the capabilities of other types of speech features. Moreover, it provides the possibility of evaluating the relative suitability of different speech feature candidates for a specific task prior to building the whole application.

6 References

1. Campbell, J.: Speaker Recognition: A tutorial. *Proceedings of the IEEE*, 85, Issue 9, 1437 – 1462 (1997).
2. Bimbot, F, et. al.: An overview of the CAVE project research activities in speaker verification. *Speech Communication*, 31, Issues 2-3, 155--180 (2000).
3. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28, Issue 4 , 357 – 366 (1980).
4. Nwe, T. L., Foo, S. W., De Silva, L. C.: Speech emotion recognition using hidden Markov models. *Speech Communication*, 41, Issue 4, 603—623 (2003).
5. O'Shaughnessy, D.: *Speech Communication: Human and Machine*, Addison-Wesley(1987).
6. Lee, K.F., Hon, H.W., Reddy, R.,:An overview of the SPHINX speech recognition system. *IEEE Transactions on Signal Processing*, 38, 35 – 45 (1990).
7. Sivakumaran P.: Robust Text Dependant Speaker Verification. Ph.D. thesis, University of Hertfordshire (1998).
8. Reynolds D., Andrews W, et.al.: The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. *Proc. ICASSP 03*, 4, 784 – 784 (2003).
9. Johnson S.: Speaker Tracking. M.Phil. Thesis, C.U.E.D. - University of Cambridge (1997).
10. Umesh, S., Cohen, L., Marinovic N., Nelson D.J.: Scale transform in speech analysis. *IEEE Transactions on Speech and Audio Processing*, 7 ,Issue 1 , 40 – 45 (1999).
11. Liu, C.S, Huang, C.S., Lin, M.T., Wang H.C.: Automatic speaker recognition based upon various distances of LSP frequencies. *Proc. IEEE International Carnahan Conference on Security Technology*, 104 – 109 (1991).
12. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*. Academic Press, (1990).

Multimodal Speaker Identification Based on Text and Speech

Panagiotis Moschonas and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki,
Thessaloniki 54124, Greece

`pmoschon@csd.auth.gr, costas@aiaa.csd.auth.gr`

Abstract. This paper proposes a novel method for speaker identification based on both speech utterances and their transcribed text. The transcribed text of each speaker’s utterance is processed by the probabilistic latent semantic indexing (PLSI) that offers a powerful means to model each speaker’s vocabulary employing a number of hidden topics, which are closely related to his/her identity, function, or expertise. Mel-frequency cepstral coefficients (MFCCs) are extracted from each speech frame and their dynamic range is quantized to a number of predefined bins in order to compute MFCC local histograms for each speech utterance, that is time-aligned with the transcribed text. Two identity scores are independently computed by the PLSI applied first to the text and the nearest neighbor classifier applied next to the local MFCC histograms. It is demonstrated that a convex combination of the two scores is more accurate than the individual scores on speaker identification experiments conducted on broadcast news of the RT-03 MDE Training Data Text and Annotations corpus distributed by the Linguistic Data Consortium.

Key words: multimodal speaker identification, text, speech, probabilistic latent semantic indexing, Mel-frequency cepstral coefficients, nearest neighbor classifier, convex combination

1 Introduction

Speaker identification systems resort mainly to speech processing. Undoubtedly, speech is probably the most natural modality to identify a speaker [1]. Historically in speaker recognition technology R&D, effort has been devoted to characterizing the statistics of a speaker’s amplitude spectrum. Although, dynamic information (e.g., difference spectra) has been taken into consideration as well as static information, the focus has been on spectral rather than temporal characterization. The usage of certain words and phrases [2] as well as intonation, stress, and timing [3], constitute longer term speech patterns, which define “familiar-speaker” differences, a promising but radical departure from mainstream speaker recognition technology.

In this paper, we explore text that is rarely combined with speech for biometric person identification. More specifically, text refers to the time-aligned

transcribed speech that appears as rich annotation of speakers' utterances. The annotation process could be an automatic, a semi-automatic, or a manual task as is frequently the case. In the proposed algorithm, we assume that we know the start time and end time of each word in a speaker's utterance as well as its speech to text transcription. Although there are a few past works where text was exploited for speaker identification, e.g. the idiolectal differences as quantified by N -gram language models [2], to the best of authors' knowledge no multimodal approach that exploits speech and text has been proposed so far.

The motivation for building multimodal biometric systems is that systems based on a single-modality, e.g. speech, are far from being error-free, especially under noisy operating conditions. The use of complementary modalities, such as visual speech, speaker's face, yields a more reliable identification accuracy. However, the additional modalities may also be unstable due to dependence on recording conditions, such as changes in pose and lighting conditions. Text and language models, if available, do not suffer from such shortcomings.

The transcribed text of each speaker's utterance is processed by the probabilistic latent semantic indexing (PLSI)[4] that offers a powerful means to model each speaker's vocabulary employing a number of hidden topics, which are closely related to his/her identity, function, or expertise. Mel-frequency cepstral coefficients (MFCCs) are extracted from each speech frame and their dynamic range is quantized to a number of predefined bins in order to compute MFCC local histograms for each speech utterance, that is time-aligned with the transcribed text. Two identity scores are independently computed by the PLSI applied first to the text and the nearest neighbor classifier applied next to the local MFCC histograms. It is demonstrated that a late fusion of the two scores by a convex combination is more accurate than the individual scores on closed-set speaker identification experiments conducted on broadcast news of the RT-03 MDE Training Data Text and Annotations corpus distributed by the Linguistic Data Consortium [6].

The outline of the paper is as follows. In Section 2, a novel method to combine audio and text data in a single representation array is described. Speaker identification algorithms based on either text or speech are described in Section 3. Experimental results are demonstrated in Section 4, and conclusions are drawn in Section 5.

2 Biometric Data Representation

In this Section, we propose a novel representation of speaker biometric data that will be used as an input to the identification algorithms to be described in the next section. As far as text data are concerned, two sets are identified, namely the set of speaker identities and the domain vocabulary. The latter is the union of all vocabularies used by the speakers. A closed set of speaker identities S of cardinality n is assumed, i.e.

$$S = \{s_1, s_2, \dots, s_n\}. \quad (1)$$

Let W be the domain vocabulary of cardinality m :

$$W = \{w_1, w_2, \dots, w_m\}. \quad (2)$$

A two dimensional matrix \mathbf{K} whose rows refer to spoken words in W and its columns refer to the speaker identities in S is created. Its (i, j) -th element, $k_{i,j}$, is equal to the number of times the word w_i is uttered by the speaker s_j :

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & k_{1,2} & \dots & k_{1,n} \\ k_{2,1} & k_{2,2} & \dots & k_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m,1} & k_{m,2} & \dots & k_{m,n} \end{bmatrix}. \quad (3)$$

It is obvious that the “word-by-speaker” matrix \mathbf{K} plays the same role with the “term-by-document” matrix in PLSI. The only difference is that the columns are associated to speakers and not to documents. Such a representation can be modeled in terms to latent variables, which refer to topics. The models can easily be derived by applying PLSI to \mathbf{K} . To minimize the vocabulary size, one may apply stemming or some sort of word clustering. Function words (e.g. articles, propositions) are frequently rejected as well.

Next, time-aligned audio information is associated with each element of the “word-by-speaker” matrix. This is done by extracting the MFCCs [5] for each frame within the speech utterance of each spoken word. Since, the same word might have been spoken by the same speaker more than once, we should aggregate the MFCC information from multiple instances of the same word. This is done as follows.

1. For each frame within each word utterance, extract 13 MFCCs. That is, 13 MFCC sets of variable length are obtained depending on the duration of each word utterance.
2. Create the histogram of each MFCC by splitting its dynamic range into b bins. Since we do not know a priori the dynamic range of each MFCC, we need to determine the minimum and maximum value for each MFCC.
3. Finally, add the MFCC histograms for all word utterances spoken by each speaker.

Accordingly, we obtain a $13 \times b$ matrix, where b is the number of histogram bins. Let the maximum and minimum value of each MFCC be \max_c and \min_c , respectively, $c = 1, 2, \dots, 13$. The size of each bin δb_c is given by

$$\delta b_c = \frac{\max_c - \min_c}{b}, \quad c = 1, 2, \dots, 13. \quad (4)$$

Let

$$\mathbf{A}_{i,j} = \begin{bmatrix} \alpha_{1,1;i,j} & \alpha_{1,2;i,j} & \dots & \alpha_{1,b;i,j} \\ \alpha_{2,1;i,j} & \alpha_{2,2;i,j} & \dots & \alpha_{2,b;i,j} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{13,1;i,j} & \alpha_{13,2;i,j} & \dots & \alpha_{13,b;i,j} \end{bmatrix} \quad (5)$$

be the $13 \times b$ matrix whose element $\alpha_{c,t;i,j}$ denotes how many times the c -th MFCC is fallen into the t -th bin of the histogram for the i -th word spoken by the j -th speaker. It is proposed each element of the “word-by-speaker” matrix to index the pair $(k_{i,j}, \mathbf{A}_{i,j})$. If $k_{i,j} = 0$, then $\mathbf{A}_{i,j} = \mathbf{0}$. Consequently, Eq. (3) is rewritten as

$$\mathbf{K} = \begin{bmatrix} (k_{1,1}, \mathbf{A}_{1,1}) & (k_{1,2}, \mathbf{A}_{1,2}) & \dots & (k_{1,n}, \mathbf{A}_{1,n}) \\ (k_{2,1}, \mathbf{A}_{2,1}) & (k_{2,2}, \mathbf{A}_{2,2}) & \dots & (k_{2,n}, \mathbf{A}_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ (k_{m,1}, \mathbf{A}_{m,1}) & (k_{m,2}, \mathbf{A}_{m,2}) & \dots & (k_{m,n}, \mathbf{A}_{m,n}) \end{bmatrix}. \quad (6)$$

The main advantage of the proposed multimodal biometric representation is that it can easily be updated when new data arrive. When a new word or a new speaker is added (e.g. during training), one has to add a new row or column in \mathbf{K} , respectively. Another main characteristic of the data representation is that contains only integers. This has a positive impact in data storage, since in most cases, an unsigned integer needs 32 bits, whereas a double number needs 64bits [6].

3 Multimodal Speaker Identification

Having defined the biometric data representation, let us assume that the training data form the composite matrix \mathbf{K} as in Eq. (6). Let the test data contain instances of speech and text information disjoint from a speaker $s_x \in S$ whose identity is to be determined. The test data are represented by the following composite vector \mathbf{k}_x , i.e.

$$\mathbf{k}_x = \begin{bmatrix} (k_{1,x}, \mathbf{A}_{1,x}) \\ (k_{2,x}, \mathbf{A}_{2,x}) \\ \vdots \\ (k_{m,x}, \mathbf{A}_{m,x}) \end{bmatrix}. \quad (7)$$

The composite matrix \mathbf{K} and the composite vector \mathbf{k}_x must have the same number of rows, thus the domain vocabulary should be the same. By denoting the vocabulary that is used by the test speaker as W_x , we could use the union of both training and test vocabulary:

$$W_{all} = W \cup W_x. \quad (8)$$

Accordingly, new rows might be inserted to both \mathbf{K} and \mathbf{k}_x and be rearranged so that each row is associated to the same word in the domain vocabulary. The next step is to combine the training and test data in one matrix as follows:

$$\mathbf{K}_{all} = [\mathbf{K} \mid \mathbf{k}_x]. \quad (9)$$

Having gathered all the data in the unified structure, \mathbf{K}_{all} , first PLSI is applied to its $k_{i,j}$ entries in order to reveal a distribution of topics related to the textual

content uttered by each speaker in S . In the following, the topics are defined by the latent discrete random variable z that admits q values in the set

$$Z = \{z_1, z_2, \dots, z_q\} \quad (10)$$

as in [4]. Let us denote by $P(s, z)$ the joint probability that speaker s speaks about topic z . Obviously,

$$P(s, z) = P(s|z) P(z), \quad (11)$$

where $P(s|z)$ is the conditional probability of a speaker given a topic and $P(z)$ is the probability of topic. By applying the PLSI algorithm, one can estimate the constituents of Eq. (11). The expectation step of the Expectation-Maximization algorithm (EM) in PLSI yields

$$P(z|w, s) = \frac{P(z) P(w|z) P(s|z)}{\sum_{z'} P(z') P(w|z') P(s|z')}. \quad (12)$$

The maximization step is described by the following set of equations:

$$P(w|z) = \frac{\sum_s k_{w,s} P(z|w, s)}{\sum_{w',s} k_{w',s} P(z|w', s)} \quad (13)$$

$$P(s|z) = \frac{\sum_w k_{w,s} P(z|w, s)}{\sum_{w,s'} k_{w,s'} P(z|w, s')} \quad (14)$$

$$P(z) = \frac{1}{R} \sum_{w,s} k_{w,s} P(z|w, s) \quad (15)$$

where $R \equiv \sum_{w,s} k_{w,s}$. The number of iterations of the EM algorithm can be preset by the user or can be determined by monitoring a convergence criterion, such as to observe insignificant changes of the model probabilities of PLSI. A random initialization of the model probabilities is frequently applied. The number of topics is also predetermined by the user.

Let the joint probability speaker $s_j \in S$ from the training set speaks about topic z_t be

$$P_{j,t} = P(s_j, z_t), \quad 1 \leq j \leq n, \quad 1 \leq t \leq q. \quad (16)$$

Similarly, let $P_{x,t} = P(s_x, z_t)$ be the same joint probability for the test speaker s_x . Then, we can define a distance between the speakers s_x and s_j based on text information as

$$d_{PLSI}(x, j) = \frac{1}{q} \sum_{t=1}^q |P_{j,t} - P_{x,t}| \quad j = 1, 2, \dots, n \quad (17)$$

or

$$d_{PLSI}(x, j) = \frac{1}{q} \sum_{t=1}^q P_{j,t} \log \frac{P_{j,t}}{P_{x,t}} \quad j = 1, 2, \dots, n. \quad (18)$$

Eq. (17) defines an L_1 -norm, whereas Eq. (18) is the KullbackLeibler divergence of the joint probabilities of speakers and topics. By applying either distance, we can obtain a vector containing all distances between the test speaker s_x and all speakers $s_j \in S$:

$$\mathbf{D}_{PLSI}(x) = [d_{PLSI}(x, 1) \ d_{PLSI}(x, 2) \ \dots \ d_{PLSI}(x, n)]^T \quad (19)$$

Let us now consider the definition of distances between speakers when local histograms of MFCCs are employed. First, we create the set of word indices L_j for each column of \mathbf{K} (i.e., the training set):

$$L_j = \{i \mid k_{i,j} > 0\}, \quad j = 1, 2, \dots, n. \quad (20)$$

Similarly, let $L_x = \{i \mid k_{i,x} > 0\}$. A distance function between the local MFCC histograms stored in $\mathbf{A}_{i,j}$ and $\mathbf{A}_{i,x}$ can be defined as

$$d_{MFCC}(j, x) = \frac{1}{|L_j \cup L_x|} \sum_{i \in (L_j \cup L_x)} \left(\frac{1}{13b} \sum_{c_1=1}^{13} \sum_{c_2=1}^b |\alpha_{c_1, c_2; i, j} - \alpha_{c_1, c_2; i, x}| \right) \quad (21)$$

where $|L_j \cup L_x|$ is the number of common words used by speakers s_j and s_x , b denotes the chosen number of MFCC local histogram bins, and $\alpha_{c_1, c_2; i, j}$ refers to the c_2 -th bin in the local histogram of the c_1 -th MFCC at the i -th word spoken by the j -th speaker column. A vector $\mathbf{D}_{MFCC}(x)$ containing the distances between the test speaker s_x and all training speakers can be defined:

$$\mathbf{D}_{MFCC}(x) = [d_{MFCC}(1, x) \ d_{MFCC}(2, x) \ \dots \ d_{MFCC}(n, x)]^T. \quad (22)$$

The elements of the distance vector in Eq. (22) can be normalized by dividing with the maximum value admitted by the distances. A convex combination of the distance vectors can be used to combine Eq. (19) and Eq. (22):

$$\mathbf{D}(x) = \gamma \mathbf{D}_{PLSI}(x) + (1 - \gamma) \mathbf{D}_{MFCC}(x) \quad (23)$$

where the parameter $\gamma \in [0, 1]$ weighs our confidence for the text-derived distance. As $\gamma \rightarrow 0$, the identification depends more on the information extracted from speech, whereas for $\gamma \rightarrow 1$ emphasis is given to the information extracted from text.

The algorithm ends by finding the minimum element value in $\mathbf{D}(x)$, whose index refers to the speaker that best matches s_x and accordingly it is assigned to s_x , i.e.:

$$s_x = \arg \min_j [\gamma d_{PLSI}(j, x) + (1 - \gamma) d_{MFCC}(j, x)]. \quad (24)$$

4 Experimental Results

To demonstrate the proposed multimodal speaker identification algorithm, experiments are conducted on broadcast news (BN) collected within the DARPA

Efficient, Affordable, Reusable Speech-to-Text (EARS) Program in Metadata Extraction (MDE). That is, a subset of the so called RT-03 MDE Training Data Text and Annotations corpus [7] is used. BN enable to easily assess the algorithm performance, because each speaker has a specific set of topics to talk about. The BN speech data were drawn from the 1997 English Broadcast News Speech (HUB4) corpus. HUB4 stem from four distinct sources, namely the American Broadcasting Company, the National Broadcasting Company, Public Radio International and the Cable News Network. Overall, the transcripts and annotations cover approximately 20 hours of BN. In the experiments conducted, the total duration of the speech recordings exceeds 2 hours.

Two sets of experiments are conducted. Both sets contain three experiments with a varying number of speakers. Speech and text modalities are treated equally. That is, $\gamma = 0.5$ in Eq. (24). Fig. 1 shows the percentage of correctly identified speakers within the R best matches for $R = 1, 2, \dots, 20$, i.e., the so called cumulative match score versus rank curve after having performed 100 iterations and chosen 4 latent topics in PLSI as well as 10 bins for each MFCC histogram. As we can see, the algorithm produces near perfect identification for 20 speakers. Concerning the group of the 37 speakers, the results are satisfactory after the 4th rank. The more difficult case, when identification among 90 speakers is sought, reveals a poor, but acceptable performance, especially after the 7th rank.

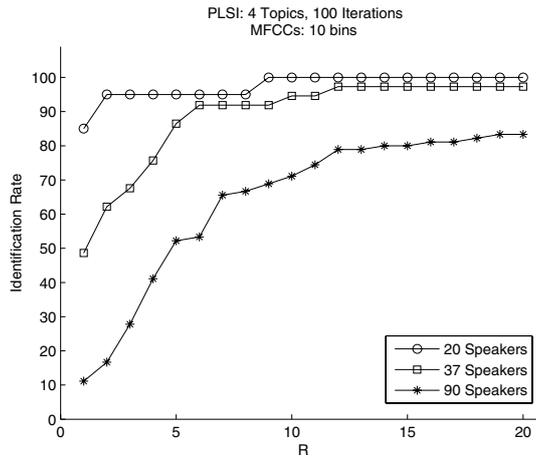


Fig. 1. Cumulative match score versus rank curve of the proposed algorithm using 4 topics and 100 iterations in PLSI model and 10 bins for every MFCC histogram.

For comparison purposes, the percentage of correctly identified speakers within the R best matches using only PLSI for the same number of iterations and topics is plotted in Figure 2. The multimodal identification offers self-evident

gains for best match identification in the case of small and medium sized speaker sets, while slight improvements of 3.32% are measured for the large speaker set.

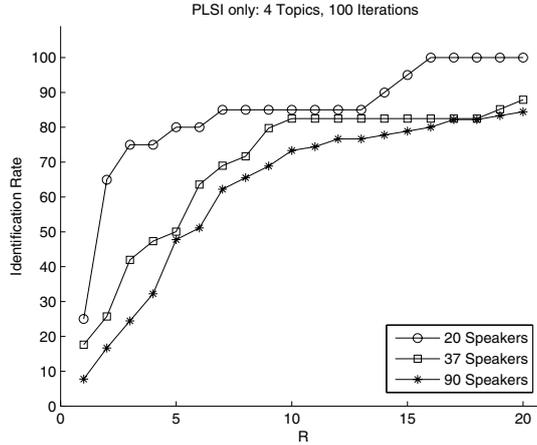


Fig. 2. Cumulative match score versus rank curve of PLSI using 4 topics and 100 iterations.

In the second set of experiments, the proposed identification algorithm is fine tuned by increasing the number of iterations to 250, the number of topics to 12, and the number of histogram bins to 50. Although, such an increase has a negative impact on the speed of the algorithm, the results are improved considerably in some cases. From the comparison of Figures 1 and 3 it is seen that the identification rate for 20 speakers is slightly increased for the best match. For the medium-sized group of 37 speakers, the identification rate for the best match is climbed at nearly 70% from 50% in the previous set. For the large group of 90 speakers, the identification rate for the best match remains the same.

By repeating the identification using only PLSI with 12 topics and 250 iterations, the percentage of correctly identified speakers within the R best matches shown in Figure 4 is obtained. The comparison of Figures 3 and 4 validates that the identification rate at best match using both text and speech increases considerably for small and medium sized speaker sets, while marginal gains are obtained for large speaker sets. Moreover, the increased number of latent topics and iterations in PLSI have helped PLSI to improve its identification rate.

5 Conclusions

In this paper, first promising speaker identification rates have been reported by combining in a late fusion scheme text-based and speech-based distances in

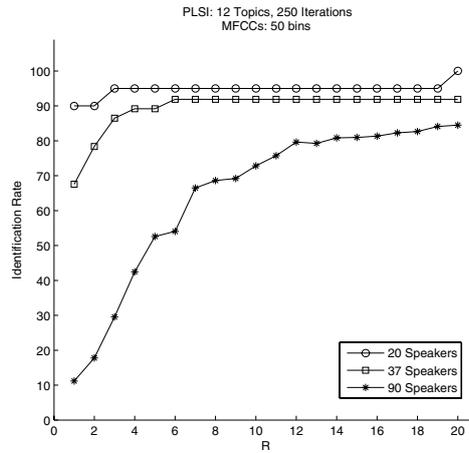


Fig. 3. Cumulative match score versus rank curve of the proposed algorithm using 12 topics and 250 iterations in PLSI model and 50 bins for every MFCC histogram.

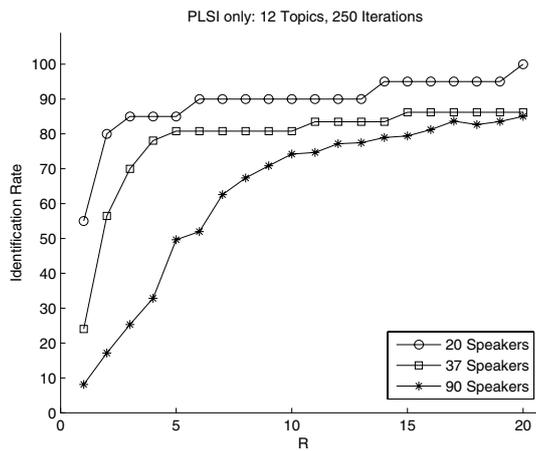


Fig. 4. Cumulative match score versus rank curve of PLSI only using 12 topics and 250 iterations.

experiments conducted on broadcast news of the RT-03 MDE Training Data Text and Annotations corpus. Motivated by the promising results, we plan to integrate MFCC histograms and document word histograms in PLSI, since both features are of the same nature and to study their early fusion.

Acknowledgments. This work has been performed within the COST Action 2101 on Biometrics for Identity Documents and Smart Cards.

References

1. Campbell, P. J.: "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462 (1997).
2. Doddington, G.: "Speaker recognition based on idiolectal differences between speakers," In Proc. *Eurospeech*, pp. 2521-2524 (2001).
3. Weber, F., Manganaro, L., Peskin, B., and Shriberg, E.: "Using prosodic and lexical information for speaker identification," In Proc. *2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*.
4. Hofmann, T.: "Probabilistic latent semantic indexing.," In: Proc. *22nd Annual Int. Conf. Research and Development in Information Retrieval (SIGIR-99)*, pp. 50-57 (1999).
5. Davies, S. B. and Mermelstein, P.: "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions Acoustic, Speech, and Signal Processing*, vol. 31, pp. 793-807, (1983).
6. Schildt, H.: *C++ The Complete Reference*, 4/e. Osborne/McGraw-Hill, N. Y. (2002).
7. Strassel, S., Walker, C., and Lee H.: RT-03 MDE Training Data Text and Annotations, Linguistic Data Consortium (LDC), Philadelphia (2004).

A Palmprint Verification System Based on Phase Congruency Features

Vitomir Štruc and Nikola Pavešić

Faculty of Electrical Engineering,
University of Ljubljana,
Tržaška 25, SI-1000 Ljubljana, Slovenia
{vitomir.struc,nikola.pavesic}@fe.uni-lj.si.com
<http://luks.fe.uni-lj.si/>

Abstract. The paper presents a fully automatic palmprint verification system which uses 2D phase congruency to extract line features from a palmprint image and subsequently performs linear discriminant analysis on the computed line features to represent them in a more compact manner. The system was trained and tested on a database of 200 people (2000 hand images) and achieved a false acceptance rate (FAR) of 0.26% and a false rejection rate (FRR) of 1.39% in the best performing verification experiment. In a comparison, where in addition to the proposed system, three popular palmprint recognition techniques were tested for their verification accuracy, the proposed system performed the best.

Key words: Palmprint verification, 2D phase congruency, Linear discriminant analysis

1 Introduction

Biometrics is a scientific discipline that involves methods of automatically recognizing (verifying or identifying) people by their physical and/or behavioral characteristics. Many biometric systems have already been presented in the literature, among them, systems which exploit biometric traits such as fingerprints, face, voice, iris, retina, hand-geometry, signature or palmprints are the most common [1].

Each of the listed biometric characteristics has its own strengths and weaknesses and is consequently more or less suited for a particular application domain. Face- and voice-based recognition systems, for example, are considered to be unintrusive, they do, however, still have problems achieving high recognition accuracy, especially when biometric samples (i.e., face images or speaker recordings) are captured in uncontrolled environments. Iris and retinal recognition, on the other hand, exhibit high recognition accuracy, but require intrusive acquisition systems [2]. Opposed to these recognition systems, palmprint-based recognition is considered both user-friendly as well as fairly accurate and thus provides an attractive alternative to other biometric systems.

Existing (unimodal) palmprint recognition systems can according to [3] (based on the employed feature extraction technique) be classified into one of three groups: texture-based (e.g., [4]), line-based (e.g., [5, 6]) and appearance-based (e.g., [7, 8]). Though all feature types are relevant for palmprint-based biometric recognition, this paper focuses on line-based features.

Most of the palmprint recognition systems that make use of line features to verify the identity of a user employ gradient-based methods to extract characteristic lines from a palmprint image (e.g., [7, 8]). While these methods work fine on images of an appropriate quality (e.g., acquired in controlled illumination condition, free of distortions caused by the pressure applied to the surface of the scanner, etc.), they have problems when features have to be extracted from palmprint images of a poorer quality. In these situations a more robust approach is preferable. To this end, we have developed a palmprint verification system that uses line features extracted with the phase congruency model and is therefore relatively insensitive to image distortions caused by the acquisition procedure (note that images acquired with a desktop scanner almost always contain regions distorted by pressure - see Fig. 1).

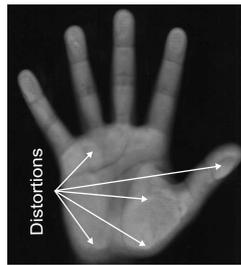


Fig. 1. Distortions of a palmprint image acquired with a desktop scanner

The rest of the paper is organized as follows: Section 2 gives a short description of the proposed palmprint verification system; Section 3 describes a series of verification experiments and presents their results; Section 4 concludes the paper with some final remarks and directions for future work.

2 System Description

The block diagram of the proposed palmprint recognition system is shown in Fig. 2. It is comprised of the following five modules: an acquisition module which uses a desktop scanner to capture an image of the palmar surface of the hand; a pre-processing module that extracts the region of interest (ROI), i.e., the palmprint region, from the acquired image and normalizes the extracted ROI in respect to size, rotation and illumination; a feature-extraction module which computes a set of phase congruency (PC) features from the normalized palmprint image and subsequently performs the linear discriminant analysis (LDA) on the feature

set to enhance its discriminatory power; a matching module that compares the computed feature set with a template (i.e., the mathematical representation of the feature sets extracted during the enrollment session) and outputs a matching score; and a decision module that uses the matching score to decide whether the person presented to the system is who he/she claims to be. A detailed description of each of the listed modules is given in the remainder of this section.

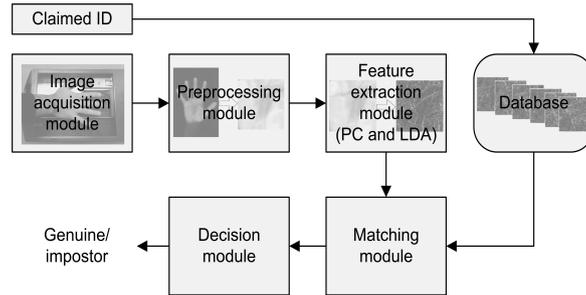


Fig. 2. The block diagram of the proposed palmprint recognition system

2.1 Image acquisition

The image-acquisition module of the proposed palmprint recognition system records grey-scale images of the palmar surface of the hand with the help of an optical desktop scanner rated at a resolution of 180 dots per inch (256 grey levels). When a person is presented to the system, he/she simply positions his/her hand on the scanner with the fingers spread naturally [1]. The system then acquires an image of the hand and passes it on to the preprocessing module.

2.2 Image preprocessing

After the acquisition stage, the acquired hand image is subjected to the preprocessing procedure which employs the following steps to extract and normalize the palmprint ROI from the hand image:

- *Binarization*: In the first step the hand region is extracted from the acquired grey-scale hand image (Fig. 3a) using an image thresholding procedure. Since a desktop scanner is employed in the acquisition stage the background of the image always appears as a black area in the image and the same (global) threshold can be used for binarization of all hand images (Fig. 3b).
- *Contour extraction*: In the second step the contour of the hand is extracted from the binarized hand image and used as the foundation for the palmprint localization procedure (an example of the extracted contour is shown in Fig. 3).

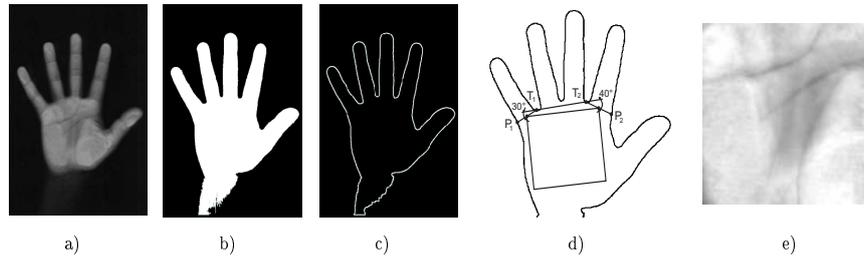


Fig. 3. The preprocessing procedure: a) The hand image acquired by the desktop scanner, b) The binary image of the hand region, c) The image of the contour of the hand region, d) Extraction of the palmprint ROI, e) The normalized palmprint image

- *ROI localization:* To locate the palmprint ROI in the hand image, two reference points are determined in the third step of the preprocessing procedure. The first, denoted as T_1 in Fig. 3d, is located at the local minimum of the hand contour between the little and the ring finger, while the second, denoted as T_2 in Fig. 3d, is set at the local minimum of the contour between the index and the middle finger. Based on the line connecting the reference locations T_1 and T_2 two additional points, i.e., P_1 and P_2 , are determined on the hand contour as shown in Fig 3d. Finally, the palmprint ROI is located as the square region whose upper two corners correspond to the middle points of the line segments P_1-T_1 and T_2-P_2 [1, 9].
- *Normalization:* In the last step the final palmprint ROI is obtained by rotating the cropped palmprint region to a predefined orientation and resizing it to a fixed size of 64×64 pixels. The geometrically normalized sub-image is ultimately subjected to an illumination normalization procedure which removes the mean of the pixel values from the grey-scale sub-image and subsequently scales all pixels with their standard deviation. An example of the normalized palmprint region is shown in Fig. 3e.

2.3 Feature extraction

The feature vector used in the matching procedure of the proposed system is extracted from the normalized palmprint image in two consecutive steps: in the first step, a set of 512 phase congruency features is computed from the input image and in the second step LDA is applied on this feature set to represent the phase congruency features in a discriminative and compact manner.

Phase congruency features. There have been a number of palmprint recognition systems presented in the literature that make use of line-based features, e.g., [5, 6]. Typically, these systems use line detectors which scan the palmprint image for points of high intensity gradients to extract the line features. However, varying illumination conditions during the image acquisition stage (when images are captured with a camera-based sensor) or deformations of the palmprint region caused by pressure applied to the surface of the scanner (when images are

captured with an optical scanner) often result in the detection of spurious lines. To avoid the listed difficulties, our systems employs the phase congruency model for line feature extraction.

The model searches for points in the palmprint sub-image where the 2D log-Gabor filter responses (of the sub-image) over several scales and orientations are maximally in phase [10, 11].

Let $\mathcal{G} = \{G(f_h, \theta_g) : h = 1, 2, \dots, p; g = 1, 2, \dots, r\}$ denote the set of 2D log-Gabor filters with p scales and r orientations and let $G(f_h, \theta_g) = G_{hg}$ be defined as:

$$G_{hg} = \exp\left\{\frac{-[\ln(f/f_h)]^2}{2[\ln(k/f_h)]^2}\right\} \exp\left\{\frac{-(\theta - \theta_g)^2}{2\sigma_\theta^2}\right\}, \quad (1)$$

where f and θ denote the polar coordinates of the log-Gabor filter in the frequency domain, f_h denotes the filters center frequency (in our experiments it was set to $f_h = 0.33 \cdot (2.1)^{1-h}$), k defines the bandwidth of the filter in the radial direction (the ratio k/f_h is commonly set to a constant value, for example, 0.55 like it was done in our case), $\theta_g = (g - 1) \cdot \pi/r$ represents the orientation of the filter and σ_θ controls the angular bandwidth of the 2D log-Gabor filter (we used a value of $\sigma_\theta = 1.2 \cdot (\pi/r)$).

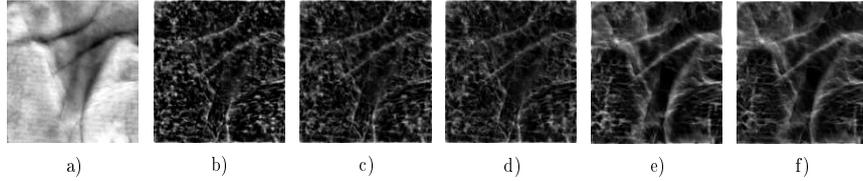


Fig. 4. a) The normalized palmprint image; Phase congruency image for b) $p = 3$ and $r = 4$, c) $p = 3$ and $r = 6$, d) $p = 3$ and $r = 8$, e) $p = 5$ and $r = 6$, f) $p = 5$ and $r = 8$

Furthermore, let $I(x)$, where x stands for the pixel location in the spatial domain, denote the grey-scale distribution of the normalized palmprint image (e.g., Fig. 3e). The magnitude $A_{hg}(x)$ and phase $\phi_{hg}(x)$ responses of the image $I(x)$ at a specific scale h and orientation g of the log-Gabor filter can then be computed as:

$$A_{hg}(x) = \sqrt{\text{Re}^2[I(x) * G_{hg}^s] + \text{Im}^2[I(x, y) * G_{hg}^s]}, \quad (2)$$

$$\phi_{hg}(x) = \arctan(\text{Im}[I(x) * G_{hg}^s] / \text{Re}[I(x) * G_{hg}^s]), \quad (3)$$

where $*$ denotes the convolution operator, G_{hg}^s stands for the log-Gabor filter in the spatial domain at the scale h and the orientation g and $\text{Re}[X]$ and $\text{Im}[X]$ represent the real and imaginary parts of the convolution output.

Finally, the two-dimensional phase congruency features can according to [10] be computed using the following expression:

$$PC_{2D}(x) = \frac{\sum_g \sum_h W_g(x) [A_{hg}(x) \Delta \Phi_{hg}(x) - T_g]}{\sum_g \sum_h A_{hg}(x) + \varepsilon}, \quad (4)$$

where T_g represents the estimated noise energy at orientation g , $W_g(x)$ denotes a weighting function that weights for the frequency spread, ε is a small constant which prevents divisions by zero, the symbols $\lfloor \cdot \rfloor$ denote the following operation:

$$\lfloor X - T \rfloor = \begin{cases} X - T, & \text{if } X > T \\ 0, & \text{otherwise,} \end{cases} \quad (5)$$

and $\Delta\Phi_{hg}(x)$ is a phase deviation measure defined as:

$$\Delta\Phi_{hg}(x) = \cos(\phi_{hg}(x) - \bar{\phi}_g(x)) - |\sin(\phi_{hg}(x) - \bar{\phi}_g(x))|. \quad (6)$$

In equation (6) $\phi_{hg}(x)$ denotes the phase angle at the location x of the log-Gabor filter phase response at scale h and orientation g , while $\bar{\phi}_g(x)$ represents the mean phase angle at the orientation g .

As we can see from the above discussion, phase congruency features are computed over multiple scales and orientation (using all filters from \mathcal{G}) making the feature extraction procedure robust to noise, illumination variations and image contrast. In addition to its robustness, the presented model also successfully explains the human perception of line (or edge) features [10].

Once a hand image is acquired, the palmprint sub-image extracted, properly normalized and transformed using the described phase congruency model, the final feature vector \mathbf{x} is constructed by dividing the phase congruency image into a number non-overlapping blocks of size 4×4 pixels and then computing the mean value and standard deviation of the pixels in each of the 256 blocks (recall that we used palmprint images of size 64×64 pixels), i.e.,

$$\mathbf{x} = (\mu_1, \sigma_1, \mu_2, \sigma_2, \dots, \mu_{256}, \sigma_{256})^T. \quad (7)$$

However, as we can see from Fig. 4, the line features extracted with the phase congruency model vary in their appearance when log-Gabor filters with different numbers of scales and orientations are used. The effects of these parameters on the verification performance of the proposed system will be evaluated in Section 3.2.

Linear discriminant analysis. Let us consider a set of n d -dimensional training phase congruency feature vectors \mathbf{x}_i arranged in a $d \times n$ column data matrix \mathbf{X} , i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ and let us assume that each of the feature vectors belongs to one of C classes (i.e., subjects - clients of the system). Based on the training data contained in the matrix \mathbf{X} , LDA first identifies a subspace (i.e., a subspace projection matrix \mathbf{W}) by maximizing a class separability criterion in the form of the ratio of the between-class to the within-class scatter matrix and then projects the phase congruency feature vectors into this subspace. The class separability criterion (sometimes called Fisher's discriminant criterion) is defined as follows [7]:

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_B \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_W \mathbf{W}|}, \quad (8)$$

where \mathbf{S}_B and \mathbf{S}_W denote the between-class and within-class scatter matrices defined as:

$$\mathbf{S}_B = \sum_{i=1}^C n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \quad (9)$$

$$\mathbf{S}_W = \sum_{i=1}^C \sum_{\mathbf{x}_j \in C_i} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T, \quad (10)$$

and the symbols $\boldsymbol{\mu}$, $\boldsymbol{\mu}_i$, n_i and C_i represent the global mean of all training feature vectors, the mean vector of the training feature vectors from the i -th class, the number of feature vectors in the i -th class and the label of the i -th class respectively.

It can be shown that the LDA transformation matrix \mathbf{W} consists of the eigenvectors corresponding to the first $m \leq C - 1$ largest eigenvalues of the following eigenproblem:

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{w}_i, \quad i = 1, 2, \dots, m \quad (11)$$

Using the calculated transformation matrix $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m]$ an arbitrary phase congruency feature vector \mathbf{x} can be projected into the LDA subspace with the help of the following expression:

$$\mathbf{y} = \mathbf{W}^T \mathbf{x}. \quad (12)$$

However, in the field of palmprint recognition the number of training samples (i.e., training phase congruency feature vectors) per class is usually significantly smaller than the number of elements contained in each of the samples. This fact makes the matrix \mathbf{S}_W singular (its' rank is at most $n - C$) and the computation of the transformation matrix \mathbf{W} using equation (11) impossible. To overcome this problem, we first projected the matrices \mathbf{S}_B and \mathbf{S}_W into the principal component subspace to ensure that the matrix \mathbf{S}_W is nonsingular and then performed LDA in this subspace. A detailed description of the employed approach can be found in [7].

2.4 Matching and decision

At the matching stage the live feature vector \mathbf{y} of a given input palmprint image computed with the help of the procedure described in the previous section is compared to the template $\bar{\mathbf{y}}_i$ associated with the claimed identity. The following similarity measure is used to produce the matching score:

$$d(\mathbf{y}, \bar{\mathbf{y}}_i) = \frac{|\mathbf{y} \bar{\mathbf{y}}_i^T|}{\sqrt{\mathbf{y} \mathbf{y}^T \bar{\mathbf{y}}_i \bar{\mathbf{y}}_i^T}}. \quad (13)$$

If the value of the normalized correlation coefficient defined by (13) is higher than the decision threshold the live feature vector and consequently the input palmprint image are recognized as genuine, otherwise they are recognized as belonging to an impostor.

3 Experiments

3.1 Database and experimental setup

The proposed palmprint verification system was tested on hand-images of 200 subjects. During the acquisition stage each of the subjects was asked to position his/her hand on the desktop scanner 10 consecutive times, resulting in a database of 2000 images.

For testing purposes the subjects were randomly split into three groups, namely, the client group (120 subjects), the evaluation impostor group (30 subjects) and the test impostor group (50 subjects). Images belonging to subjects from the client group were further divided into sets of training images (4 per subject), evaluation images (3 per subject) and test images (3 per subject). Images from the client training set were used to construct client-templates (i.e., mean feature vectors), images from the impostor as well as the client evaluation set were used to compute the decision threshold and to optimize the system parameters (i.e., number of scales and orientations of the 2D log-Gabor filters) while the remaining test sets were employed exclusively for the final performance evaluation. During this last stage each of the 3 client test images was compared to the corresponding class in the database (a total of $3 \times 120 = 360$ experiments), whereas all 10 impostor test images were compared to each of the classes in the database (a total of $10 \times 50 \times 120 = 60,000$ experiments).

Three error rates were used in our experiments to rate the accuracy of the proposed palmprint verification system: the false acceptance rate (FAR) which measures the frequency of falsely accepted impostors, the false rejection rate (FRR) which measures the frequency of falsely rejected clients and the equal error rate (ERR) that is defined as the error rate at which the FAR and FRR are equal. In addition to providing an accuracy measure for the proposed system, the ERR (obtained on the evaluation sets) was used for determining the decision threshold.

3.2 Parameter tuning

Our first set of experiments assessed the performance of the proposed palmprint verification system with respect to the number of scales and orientations of the 2D log-Gabor filters used to compute the phase congruency features. The system was tested for 5 different combinations of the values of p and r (see Section 2.3). In all cases the number of features was set to its maximal value, i.e., $m = 119$. The results of the experiments are presented in Fig. 5 and Table 1 which show the ROC curves and the values of the FAR and FRR at the ERR operating point respectively.

As we can see, varying the number of filter orientations had only a small effect on the verification performance of the proposed system. Larger differences were detected when the number of scales was changed. Furthermore we can notice that the error rates at the equal error operating point for images processed with log-Gabor filters at 3 scales and different numbers of orientations are virtually

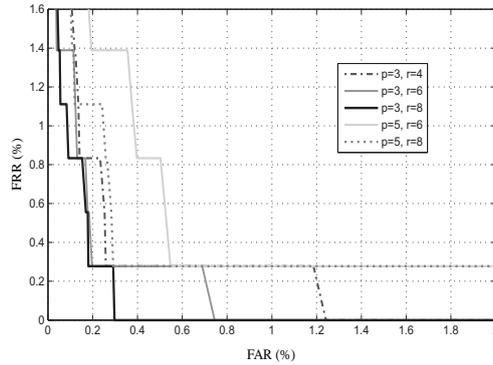


Fig. 5. The ROC curves of the performed experiments

Table 1. The FRRs and FARs of the experiments at the equal error operating point

No. of scales	No. of orient.	$FAR(\%)$	$FRR(\%)$
$p = 3$	$r = 4$	0.26	0.28
$p = 3$	$r = 6$	0.26	0.28
$p = 3$	$r = 8$	0.25	0.28
$p = 5$	$r = 6$	0.50	0.83
$p = 5$	$r = 8$	0.28	0.56

the same. However, by looking at Fig 5 we can see that the combination of 3 scales and 8 orientations performed the best (considering all possible operating points).

3.3 Performance evaluation

The goal of the second set of verification experiments was to assess the performance of the proposed system on an independent set of test images. Additionally, three popular palmprint-feature extraction techniques were implemented, trained and compared to our approach. Specifically the following methods were implemented for comparison: the eigenpalm approach [8], the fisherpalm approach [7] and a line-feature [5] based approach (denoted as LFBA in Table 2) in combination with LDA. Note, however, that the original LFBA, i.e., as presented in [11], does not use LDA to extract the final palmprint features. LDA was added to allow for a fair comparison with the proposed approach which also includes a LDA step.

The results of the experiments in terms of the FRR and FAR obtained with the threshold that ensured equal error rates on the evaluation set are presented in Table 2. Two findings should be emphasized based on these results: first, the FRRs of all methods increased in the final testing stage, most likely due to an unrepresentative training set which did not account for all possible variations in the appearance of the line features of the client images; and second, the proposed line features resulted in the best verification performance of all tested methods.

Table 2. Comparison of the FRRs and FARs for different feature extraction techniques

Feature extraction procedure	<i>FAR</i> (%)	<i>FRR</i> (%)
Eigenpalm	2.94	3.61
Fisherpalm	0.30	1.94
LFBA	0.39	2.22
Proposed approach	0.26	1.39

4 Conclusion and future work

We have presented a palmprint recognition system that used phase congruency and linear discriminant analysis to extract discriminative palmprint features. The system was tested on a database of 2000 hand images and achieved a false acceptance rate of 0.26% and a false rejection rate of 1.39% using the decision threshold that ensured equal error rates on an independent evaluation set. Based on these encouraging results, our future work will be focused on the integration of phase congruency features into a multi-modal (i.e., intra-modal) palmprint recognition system.

References

1. Pavešić, N., Ribarić, S., Ribarić, D.: Personal Authentication Using Hand-Geometry and Palmprint Features - The State of the Art. In: Proceedings of the Workshop: Biometrics - Challenges Arising from Theory to Practice, pp. 17–26. Cambridge (2004)
2. Yoruk, E., Dutagaci, H., Sankur, B.: Hand Biometrics. *Image and Vision Computing* 24(5), 483–497 (2006)
3. Kumar, A., Zhang, D.: Personal Authentication Using Multiple Palmprint Representation. *Pattern Recognition* 38(10), 1695–1704 (2005)
4. Zhang, D., Kong, W.K., You, J., Wong, M.: Online Palmprint Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(9), 1041–1050 (2003)
5. Zhang, D., Shu, W.: wo Novel Characteristics in Palmprint Verification: Datum Point Invariance and Line Feature Matching. *Pattern Recognition* 32(4), 691–702 (1999)
6. Kumar, A., Wong, D.C.M., Shen, H.C., Jain, A.K.: Personal Authentication Using Hand Images. *Pattern Recognition Letters* 27(13), 1478–1486 (2006)
7. Wu, X., Zhang, D., Wang, K.: Fisherpalms Based Palmprint Recognition. *Pattern Recognition Letters* 24(15), 2829–2838 (2003)
8. Lu, G., Zhang, D., Wang, K.: Palmprint Recognition Using Eigenpalm Features. *Pattern Recognition Letters* 24(9-10), 1463–1467 (2003)
9. Ribarić, S., Fratricić, I.: A Biometric Identification System Based on Eigenpalm and Eigenfinger Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(11), 1698–1709 (2005)
10. Kovesi, P.: Image Features From Phase Congruency. *Videre: Journal of Computer Vision Research* 1(3), 1–26 (1999)
11. Gundimada, S., Asari, V.K.: A Novel Neighborhood Defined Feature Selection on Phase Congruency Images for Recognition of Faces with Extreme Variations. *International Journal of Information Technology* 3(1), 25–31 (2006)

Some unusual experiments with PCA-based palmprint and face recognition

Ivan Krevatin and Slobodan Ribarić¹

¹ Faculty of EE and Computing, University of Zagreb, Croatia
ivan.krevatin@inet.hr, slobodan.ribaric@fer.hr

Abstract. In this paper we describe a number of experiments relating to PCA-based palmprint and face recognition. The experiments were designed to determine the influence of different training sets used for the construction of the eigenpalm and eigenface spaces on the recognition efficiency of biometric systems. The results of the recognition experiments, obtained using three palmprint databases (PolyU, FER1, FER2) and one face database (XM2VTSDB), suggest that it is possible to design a biometric recognition system that is robust enough to successfully recognize palmprints (or faces) even in cases when the eigenspaces are constructed from completely independent sets of palmprints or face images. Furthermore, the experiments show that for PCA-based face-recognition systems with an eigenspace that is constructed by using palmprint-image databases, and PCA-based palmprint-recognition systems with an eigenspace that is constructed using a face-image database, the recognition rates are unexpectedly improved compared to the classic approach.

Keywords: Biometrics, Eigenface, Eigenpalmprint, Face recognition, Palmprint recognition, Principal Component Analysis

1 Introduction

The hand and the face provide the source for a number of physiological biometric features that are used in unimodal and multimodal biometric systems for user authentication or recognition [1]–[4]. Principal component analysis (PCA) [5], also known as the Karhunen-Loeve transform, is commonly used for both palmprint [1], [6]–[8] and face recognition [9]–[11]. PCA is one of the so-called appearance-based methods, which operate directly on an image-based representation and extract features in the subspace derived from the training images. The subspace constructed using a PCA is defined by the principal components of the distribution of the training set consisting of the images of the person's palmprints (or face). This subspace is called the eigenspace.

The key point is that these images are selected from the images of all the people that will be enrolled in the system. In other words, in a biometric-based identification or verification system a PCA is used to transform the data from an original, high-

dimensional space into a feature space with significantly fewer dimensions. A PCA constructs the projection axes (which define the feature space with lower dimensionality) based on the training samples of the users taken from the original space. After that, one or more samples from the training set are projected onto these axes to obtain the feature vectors that represent the users' template(s), to be stored in a database. These templates are used in the matching process with the users' test templates (during testing of the system). The stored templates can also be treated as enrolled users' templates and they are used for matching with the users' live templates during the authentication phase.

Since the exact distribution of the palmprints (or the face images) cannot be obtained from the training samples, the projection axes are calculated based on an approximation with the limited set of training samples. It is clear that the orientations of the axes depend on how good is the approximation of the distribution. The more training samples we have (assuming that the samples are randomly selected) the better is the approximation, and the projection axes are then closer to their ideal positions. In this case ideal means the positions of the axes obtained from the exact distribution.

We based our experiments on the assumption that for a target biometric-based recognition system the distribution of the samples can be equally or even better approximated with the large number of samples that are not all obtained from the users of the target system than it can be with the limited set of samples only available during the training phase, taken from the users that will be enrolled in the system. Our intention here was to determine whether it is possible to avoid the problem of recalculating or updating the eigenspace and the stored templates for each new user that is going to be enrolled in the system.

We then extended our experiments in an unusual way, i.e., we used the set of face images for the construction of the eigenspace that is then used in the PCA-based palmprint-recognition system, and, vice versa, the set of palmprint images were used to calculate the eigenspace that is then used in the PCA-based face-recognition system.

2 Experiments and Results

The following sets of experiments were performed:

- Testing of the palmprint-recognition system with a dependent eigenspace, i.e., the system where the eigenspace was constructed from the training set of palmprint images of the users that are used for the enrolment;
- Testing of the face-recognition system with a dependent eigenspace;
- Testing of the palmprint-recognition system with an independent eigenspace, i.e., the system where the eigenspace was constructed from an independent set of palmprint images that do not belong to the users of the system;
- Testing of the palmprint-recognition system that is based on the eigenspace obtained from the training set of the face images;
- Testing of the face-recognition system that is based on the eigenspace obtained from the training set of the palmprint images.

In order to conduct the above sets of experiments, two types of biometric-based recognition systems were built: a system with a dependent eigenspace and a system with an independent eigenspace. Both types are related to palmprint- and face-based recognition. Fig. 1 shows the block diagram of the system with a dependent eigenspace; Fig. 2 shows the block diagram of the system with an independent eigenspace.

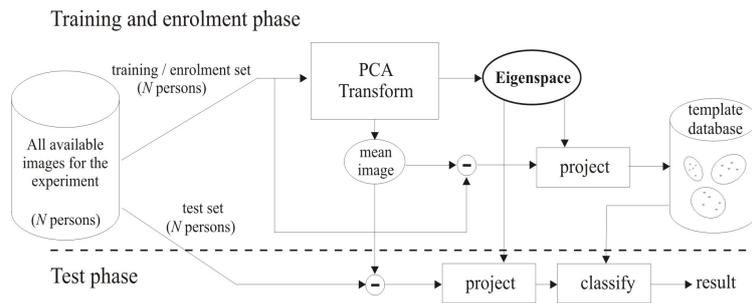


Fig. 1. Block diagram of the system with a dependent eigenspace

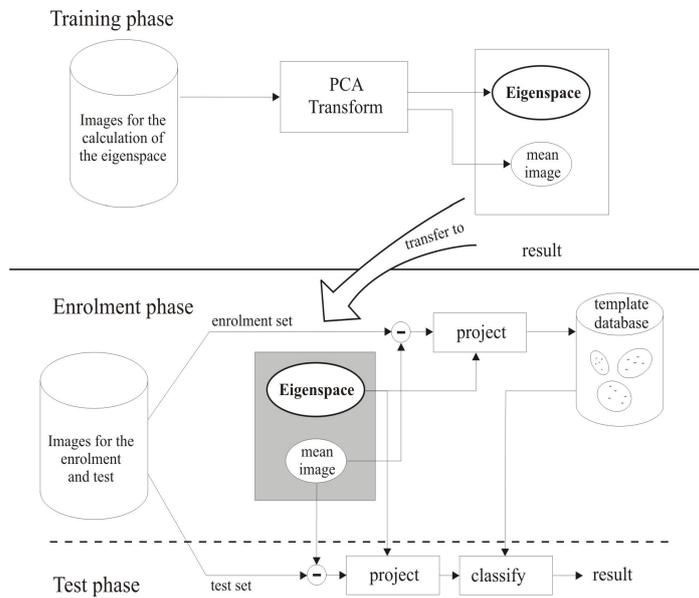


Fig. 2. Block diagram of the system with an independent eigenspace

For both types of system we used the 1-Nearest Neighbour rule as the classification method, with the Euclidean distance as a measure of the dissimilarity.

2.1 Databases

There are four basic databases that are used in our experiments: FER1, FER2, XM2VTSDB [12] and PolyU [13]. For the purpose of the experiments we collected two independent palmprint databases (FER1 and FER2) with images taken using a desktop scanner. The structure of the databases is as follows: FER1 has a total of 545 palmprint images taken from 109 people with 5 images per person and FER2 has a total of 752 images collected from 94 people with 8 images from each. The databases were created on different occasions and it was ensured that none of the people who gave their palmprints for one of the databases did the same for the other one. The images were scanned at 256 grey levels, with a resolution of 180 dpi for the FER1 database and 150 dpi for the FER2 database.

The XM2VTSDB database contains 1180 face images taken from 295 people, with 4 images from each person.

The PolyU database contains palmprint images from 386 different palms, captured with a specialized device using a camera. Because of the number of images it contains, the PolyU database is suitable for our experiments to test the systems with an independent eigenspace. For this purpose we randomly selected images from the PolyU database to form three databases to be structurally equal (the same number of people and images per person) to FER1, FER2 and XM2VTSDB databases. These databases are referred to as PolyU1, PolyU2 and PolyU3, respectively.

Table 1. shows the databases used in the experiments.

Table 1. Databases used in the experiments

Database	Number of users	Total number of images
FER1 (palmprint, 150dpi/256 grey levels)	109	545 (5 images per user)
FER2 (palmprint, 180dpi/256 grey levels)	94	752 (8 images per user)
PolyU (palmprint, CCD based capturing device)	386	7752
XM2VTSDB (face)	295	1180 (4 images per user)
PolyU1 (derived from PolyU)	109	545 (5 images per user)
PolyU2 (derived from PolyU)	94	752 (8 images per user)
PolyU3 (derived from PolyU)	295	1180 (4 images per user)

2.2 Preprocessing

From the palmprint images we extracted a square region of interest (ROI) from the centre of the palm. The ROI is defined on the basis of two stable points on the contour of the hand: the first is located in the valley between the little finger and the ring finger, and the second is located between the index finger and the middle finger. The preprocessing phase for the scanned images from the FER1 and FER2 databases and the camera images of the PolyU database can be summarized in the following steps: (i) global thresholding; (ii) border following; (iii) locating the region of interest

(ROI); (iv) extracting the ROI and compensating for its rotation; (v) applying the Gaussian mask; (vi) resizing the ROI to a size of 40 by 40 pixels; and (vii) performing a histogram equalization. Fig. 3 illustrates the phases of preprocessing for a palmar hand image from the FER1 database.

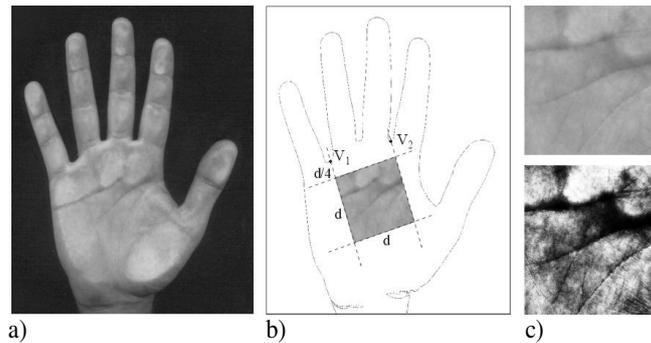


Fig. 3. Illustration of the preprocessing of a palmar hand image from the FER1 database: a) input image, b) localization of the ROI, and c) ROI (40x40 pixels) before and after the histogram equalization

The face images from the XM2VTSDB database were normalized using the normalization method described in [8]. Using this method the elliptical region of the face is detected and the background is removed. To obtain images compatible with the palm ROI images, the centre of the normalized image, with a size of 40 by 40 pixels, is extracted and a histogram equalization is performed (Fig. 4).

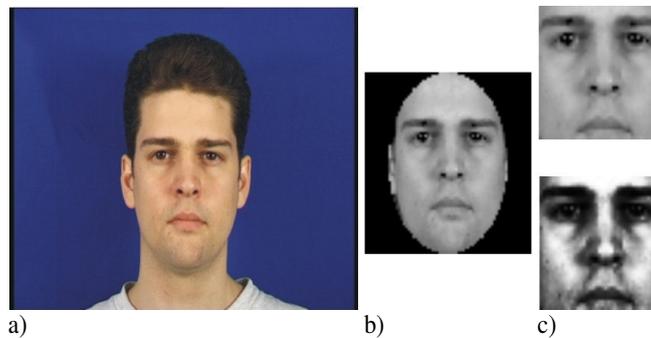


Fig. 4. Illustration of the preprocessing phases for face images: a) input image, b) normalized image, and c) cropped ROI (40x40 pixels) before and after the histogram equalization

2.3 Palmprint- and face-recognition systems with a dependent eigenspace

The experiments were performed according to the scenario described in Fig. 1. The results obtained in these experiments were used as a reference for comparing with the other systems.

The number of images per person used for the training and enrolment varied, depending on the structure of the database. In the case of the FER1 and PolyU1 databases three of five images from each person were used for the training and enrolment, and the remaining two images were used for the recognition. This means that the total number of images used for the construction of the eigenspace was $109 \times 3 = 327$, where 109 is the number of people. In the case of the FER2 and PolyU2 databases, four images per person were used for the training and enrolment, and the remaining four were used for the testing. The eigenspace was built from a set of $94 \times 4 = 376$ images, where 94 is the number of people. In the last case, for the databases XM2VTSDB and PolyU3, two images per person were used for the enrolment and training, and the remaining two were used for the testing. A total of $295 \times 2 = 590$ images were used to build the eigenspace, where 295 is the number of people.

During the enrolment phase, all the training images that were used to build the eigenspace were projected onto it to form a set of users' templates and then stored in the system database.

Recognition rates were calculated for the various lengths of the feature vectors, i.e., for the number of PCA components. In each experiment the training and testing images were randomly picked from the set of available samples. The recognition rates were averaged over 10 experiments, except for the XM2VTSDB and the PolyU3 databases where, because of the small set of samples, only six experiments were performed. The results are shown in Table 2.

Table 2. Recognition rates of the systems with a dependent eigenspace

Database		Recognition rates (%)					
		Number of PCA components					
		25	50	100	150	200	250
Palmprint	FER1	94.13	96.19	96.61	96.88	96.84	96.84
	FER2	89.87	92.29	93.32	93.51	93.56	93.59
	PolyU1	92.94	94.54	94.50	94.40	94.27	94.17
	PolyU2	93.48	94.97	95.27	95.16	95.16	95.11
	PolyU3	84.66	87.77	86.86	86.44	85.68	85.57
Face:							
	XM2VTSDB	68.98	75.80	78.31	78.68	79.09	79.22

The best palmprint-recognition rates vary from 93.59% to 96.88% for the different databases and for the different lengths of the feature vectors. The best result for the face recognition was obtained for 250 components of a feature vector (79.22%).

2.4 Palmprint-recognition systems with an independent eigenspace

The experiments with an independent eigenspace were performed by following the scenario described in Fig. 2. In each experiment one palmprint database was used for the construction of the eigenspace (the training database) and another one for the enrolment and recognition (the testing database). The two databases were chosen to be structurally equal, i.e., with the same number of images per person and the same

number of people. This means that in the case when the PolyU1 database was used for the construction of the eigenspace, the FER1 database was used for the enrolment and testing, and vice versa, when the FER1 database was used for the construction of the eigenspace, the PolyU1 database was used for the enrolment and testing. Since in the experiments with the dependent eigenspace not all the images from the database could be used for the construction of the eigenspace (some of the images were separated for the testing set) we did the same in these experiments. We first performed the tests with the eigenspace constructed from part of the database, like it was with the dependent eigenspace (for the FER1 and PolyU1 databases, 327 instead of 545 images), and then we performed the same set of experiments, but this time using the eigenspace constructed from the complete databases (all 545 images in the case of the FER1 and PolyU1 databases).

The results of the palmprint-recognition systems with an independent eigenspace are summarized in Table 3. The results of the experiments where the complete database was used for the construction of the eigenspaces are marked with (*) (Table 3.).

The selection of the enrolment and testing samples, as well as the way in which the recognition rates were calculated, was the same as described for the experiments with a dependent eigenspace. In this way the obtained results were suitable for a comparison with the results from Table 2.

Table 3. Recognition rates of the palmprint-recognition systems with an independent eigenspace

Database for eigenspace	Enrolment and testing database	Recognition rates (%)					
		Number of PCA components					
		25	50	100	150	200	250
PolyU1	FER1	89.77	93.21	95.14	95.96	96.06	96.06
PolyU1 (*)	FER1	92.02	95.64	96.74	97.02	96.97	97.11
FER1	PolyU1	94.13	95.23	95.55	95.83	95.78	95.78
FER1 (*)	PolyU1	94.26	95.60	96.15	95.55	95.78	95.50
PolyU2	FER2	85.56	90.29	92.85	93.19	93.19	93.35
PolyU2 (*)	FER2	85.77	90.43	92.87	93.30	93.51	93.78
FER2	PolyU2	93.61	95.27	95.74	95.90	95.85	95.61
FER2 (*)	PolyU2	93.27	95.24	95.77	95.82	95.66	95.53

(*) – eigenspace constructed from the complete database

The results show that the use of a larger number of images (the complete database) for the construction of the eigenspace improved the recognition rates for all the databases, except in the case when the PolyU2 database was used for the enrolment and testing and the FER2 database was used for the construction of the eigenspace. In this case the results deteriorated slightly. The improvement is particularly noticeable when the FER1 database was used for the enrolment and testing.

When comparing the results to those systems with a dependent eigenspace (Table 2.) we can see that for all the databases, the overall best recognition rates were better for the systems with an independent eigenspace (!). In the case when the PolyU1 and PolyU2 databases were used for the enrolment and testing the system with an

independent eigenspace outperformed the system with a dependent eigenspace, when a small number of PCA components (25) was used, which is not the case for the FER1 and FER2 databases. We should draw attention to the case of the PolyU1 database, where for the same number of PCA components (50) the recognition rate of the system with an independent eigenspace (95.60%) was more than 1% better than the best result of the system with a dependent eigenspace (94.54%). In the same case the best result of the independent eigenspace system was 96.15%, which was achieved for 100 PCA components.

In the worst case for the system with an independent eigenspace, which happened when the FER1 database was used for the enrolment and testing, the difference between the best recognition rates was less than 1% (96.88% compared to 96.06%), in favour of the system with a dependent eigenspace. The overall best recognition rate (97.11%) was achieved for the same database in the system with an independent eigenspace, where the whole set of images from the PolyU1 database was used for the construction of the eigenspace. The above results support our assumption outlined in the Introduction.

2.5 Palmprint recognition using a face eigenspace and face recognition using a palm eigenspace

We extended, in a slightly unusual way, the experiments with the independent eigenspace in such a manner as to use the eigenspace constructed from the face images in the palmprint-recognition system, and vice versa, the eigenspace constructed from palmprint images in the face-recognition system. For this purpose we used the XM2VTSDB face database and the structurally equal PolyU3 palmprint database (see Table 1.).

We first performed the palmprint-recognition experiments using only two face images per person from the database XM2VTSDB for the construction of the eigenspace, as was the case in the system with a dependent eigenspace, and then using all the face images from the database XM2VTSDB.

After that, in a similar way, we performed the face-recognition experiments using only two palmprint images per person from the database PolyU3, and then using the complete database PolyU3 for the construction of the eigenspaces.

The results of the palmprint- and face-recognition are shown in Table 4. The recognition rates are averaged over six experiments. The experiments where all the images of the database were used for the construction of the eigenspace are marked with (*).

For the palmprint-recognition systems the best recognition rate was improved from 87.77%, in the system with a dependent eigenspace (for PolyU3 database, Table 2.), to 89.38%, in the system with an eigenspace constructed from the face images (Table 4.) (!). In the latter the best recognition rate was achieved when using a larger number of PCA components, i.e., 150, in comparison to the dependent system, where the best recognition rate was for 50 PCA components.

Use of a larger set of palmprint images for the construction of the eigenspace improved the face-recognition results compared to the case when the XM2VTSDB face database was used for the enrolment and testing. The best result for the face

recognition improved from 79.22%, in the system with a dependent eigenspace (see Table 2.), to 80.14%, in the system that used the eigenspace constructed from the palmprint images of the PolyU3 database, and to 81.10% when the complete PolyU3 database (*) was used. In both face-recognition systems the best results were achieved using the same number of PCA components (250).

Table 4. Recognition rates of the palmprint- and face-recognition systems

		Recognition rates (%)					
Database for eigenspace	Enrolment and testing database	Number of PCA components					
		25	50	100	150	200	250
XM2VTSDB	PolyU3	80.68	87.29	88.98	89.38	89.12	88.92
XM2VTSDB (*)	PolyU3	80.90	87.09	88.64	88.70	88.59	88.36
PolyU3	XM2VTSDB	59.35	70.81	77.51	79.35	79.83	80.14
PolyU3 (*)	XM2VTSDB	59.57	71.07	78.84	80.02	80.56	81.10

(*) – eigenspace constructed from the complete database

3 Conclusions

In this paper we describe a number of experiments with PCA-based palmprint and face recognition. The experiments were designed to determine the influence of the different training sets used for the construction of the eigenpalm and eigenface spaces on the recognition accuracy of biometric-based recognition systems. The experiments can be divided into the following sets:

- i) in the first set of experiments we performed a test of the PCA-based palmprint-recognition systems using a classic approach, where the eigenspaces were constructed from the training sets of palmprint images of the users that are used for the enrolment. The same approach was used for a PCA-based face-recognition system.
- ii) in the second set of experiments, in order to test the robustness of the systems, the eigenspaces were constructed from an independent set of palmprint images from the users that were not enrolled in the system.
- iii) in the third set of (unusual) experiments, for the PCA-based palmprint recognition we used the eigenspaces calculated from the face images, and for the PCA-based face recognition we used the eigenspaces obtained from the palmprint images.

A summary of the results is as follows:

- i) comparing the results of the palmprint recognition obtained with the PCA-based systems with a dependent eigenspace (Table 2.) with the results obtained with the PCA-based systems with an independent eigenspace (Table 3.) it is clear that for all the databases the overall best recognition rates are better for the systems with an independent eigenspace (!);
- ii) for PCA-based face-recognition systems with an eigenspace constructed using palmprint-image databases, the recognition rates were unexpectedly

improved (compared to the classic approach) from 79.22% (Table 2.) to 80.14% and 81.10% (Table 4.) for the PolyU3- and PolyU3(*)-based eigenspaces, respectively.

The results of the experiments obtained using the three basic palmprint databases (PolyU, FER1, FER2) and the three derived palmprint databases (PolyU1, PolyU2 and PolyU3), and the single face database (XM2VTSDb) (discussed in detail in Section 3) led to the interesting main conclusion that it is possible to design a biometric-based recognition system that is robust enough to successfully recognize palmprints (or faces) even in the case when the eigenspaces are constructed from completely independent sets of palmprint or face images. From this it follows that there is no need to construct a new eigenspace or apply methods for an incremental eigenspace update [14], [15] when new users are enrolled in the system. Furthermore, it will be possible to install a biometric-based PCA-authentication system with predefined projection axes that are independent of users' database for a specific application. Of course, we are aware that the above conclusions are a little rash and that they will have to be verified by further experiments, and so we are planning to test our findings on larger palmprint and face databases.

References

1. Zhang, D. D.: *Palmprint Authentication*. Kluwer Academic Publishers, Boston, (2004)
2. Maltoni, D. et al.: *Handbook of Fingerprint Recognition*. Springer, New York, (2003)
3. Li, SZ., Jain, A.K.: *Handbook of Face Recognition*. Springer, New York, (2005)
4. Kumar, A., Zhang, D.: Integrating palmprint and face for user authentication. In: *Proc. Multi Modal User Authentication Workshop*, pp. 107--112. Santa Barbara, (2003)
5. Jolliffe, I. T.: *Principal Component Analysis*. Springer, New York, (1986)
6. Lu, G., Zhang, D., Wang, K.: Palmprint recognition using eigenpalms. *Pattern Recognition Letters* 24, 1463--1467, (2003)
7. Connie, T., Teoh, A., Goh, M., Ngo, D.: Palmprint Recognition with PCA and ICA. *Conference of Image and Vision Computing New Zealand 2003*, pp. 227--232, (2003)
8. Ribarić, S., Fratrić, I.: A Biometric Identification System Based on Eigenpalm and Eigenfinger Features. *IEEE Trans. Patt. Anal. Mach. Intell.* 27, pp. 1698--1709, (2005)
9. Turk, M., Pentland, A.: Eigenfaces for Recognition. *J. of Cognitive Neuroscience* 3, pp. 71--86, (1991)
10. Belhumeur P. N., Hespanha, J. P., Kriegman, D. J.: Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE Trans. Patt. Anal. Mach. Intell.* 19, pp. 711--720, (1997)
11. Moghaddam, B.: Principal manifolds and bayesian subspaces for visual recognition. *IEEE Trans. Patt. Anal. Mach. Intell* 24, pp. 780--788, (2002)
12. XM2VTSDb Face Database, <http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb>
13. PolyU Palmprint Database, <http://www.comp.polyu.edu.hk/~biometrics/>
14. Ozawa, S., Pang, S., Kasabov, N.: A Modified Incremental Principal Component Analysis for On-line Learning of Feature Space and Classifier. *8th Pacific Rim International Conference on Artificial Intelligence*, pp. 231--240, Auckland, (2004)
15. Hall, P., Martin, R.: Incremental Eigenanalysis for Classification. In: *Proc. British Machine Vision Conference*, pp. 286--295, (1998)

An Empirical Comparison Of Individual Machine Learning Techniques In Signature And Fingerprint Classification

Márjory Abreu and Michael Fairhurst

Department of Electronics, University of Kent, Canterbury, Kent CT2 7NT, UK
{mcda2, M.C.Fairhurst}@kent.ac.uk

Abstract. This paper describes an empirical study to investigate the performance of a wide range of classifiers deployed in applications to classify biometric data. The study specifically reports results based on two different modalities, the handwritten signature and fingerprint recognition. We demonstrate quantitatively how performance is related to classifier type, and also provide a finer-grained analysis to relate performance to specific non-biometric factors in population demographics. The paper discusses the implications for individual modalities, for multiclassifier but single modality systems, and for full multibiometric solutions.

Keywords: Classifiers, signature, fingerprints.

1 Introduction

Optimising the processing of biometric identity data, whether within modalities or in multimodal form, is a fundamental challenge in system design and deployment. There are many potential options available in relation to the processing engines which might be adopted, and any selection must be made on the basis both of application requirements and with regard to a knowledge of the degree of match between the underlying population data distributions and system operating characteristics.

The availability of multiple information sources for biometric data processing can suggest various different strategies by means of which to achieve enhanced performance. These include, for example, selecting an optimal processing technique from among many options, combining processors to create a multiple processor system to work on a single modality source and, ultimately, combining multiple biometric modalities to overcome the shortcomings of any one individual modality. In each case, however, there are obvious questions to be asked about the processing engines implemented, and the performance of which they are inherently capable.

This paper reports on an empirical study which addresses these fundamental questions. Specifically, we investigate the application of a wide range of different possible techniques for the classification of biometric data. We will present performance metrics which show quantitatively how the choice of classifier will

determine the performance which can subsequently be achieved by a system operating within a specific modality. We then demonstrate how a lower-level analysis can deliver more targeted selection strategies in situations where outcome might be guided by the availability of specific information which can inform the decision-making process (the availability of demographic/non-biometric data, for example). Our investigation will also contribute to the development of approaches to the implementation of multi-classifier solutions to identification processing based on a single modality, providing performance indicators across a range of classifiers which might be adopted in such a multiple classifier configuration.

Finally, because we will present experimental data from two (fundamentally different) modalities, our study will be valuable in pointing towards some issues of relevance in multimodal processing configurations in future studies. We have chosen, on the one hand, fingerprint processing to illustrate the use of a physiological biometric of considerable current popularity and wide applicability and, on the other hand, the handwritten signature, a behavioural biometric which is currently less widely adopted, in order to give a broad base to our study and to allow the most general conclusions to be drawn.

Our study will therefore provide both some useful benchmarking for system implementation, and a logical starting point for further development of practical systems for effective and efficient biometric data processing.

2 Methods And Methodology

We report some experiments based on two biometric modalities, respectively fingerprint images and handwritten signature samples. The databases used for experimentation are described in detail in Section 3. Since the focus of our study is on the performance of different classifier types, we identify a pool of specific classification algorithms giving a broad representation of different approaches and methodologies.

In our experiments, each database is divided in two sets, one of which (containing approximately 90% of the samples) is used to train the classifier and the other of which (10%) is used to validate the method. The 10-cross-validation method [13] is used to evaluate classifier performance. In this evaluation method, the training set is divided into ten folds, each with approximately the same number of samples. Thus, a classifier is trained with nine folds and tested with the remaining unused fold. Validation is performed every time the test fold is run.

The analysis of the resulting classifier performance used the statistical t-test [15] with 95% degree of confidence. This test uses t-Student distribution to compare two independent sets. The use of this test allows us to say whether a classifier is statistically more accurate than another just by observing whether the p value is smaller than the threshold established.

The pool of classifiers selected, comprising eight specific classifiers, is first briefly described.

Multi-Layer Perceptron (MLP) [12]: MLP is a Perceptron neural network with multiple layers [18]. The output layer receives stimuli from the intermediate layer and generates a classification output. The intermediate layer extracts the features, their weights being a codification of the features presented in the input samples, and the intermediate layer allows the network to build its own representation of the problem. Here, the MLP is trained using the standard backpropagation algorithm to determine the weight values.

Radial Basis Function Neural Network (RBF) [5]: This adopts an activation function with radial basis, and can be seen as a feed forward network with three layers. The input layer uses sensory units connecting the network with its environment. The second layer executes a non-linear transformation from the input space through the output space performing the radial basis function.

Fuzzy Multi-Layer Perceptron (FMLP) [6]: This classifier incorporates fuzzy set theory into a multi-layer Perceptron framework, and results from the direct "fuzzyfication" in the network level of the MLP, in the learning level, or in both. The desired output is differently calculated when compared with the MLP, the nodes which are related with the desired output being modified during the training phase, resulting in a "fuzzy output".

Support Vector Machines (SVM) [16]: This approach embodies a functionality very different from that of more traditional classification methods and, rather than aiming to minimize the empirical risk, aims to minimize the structural risk. In other words, the SVM tries to increase the performance when trained with known data based on the probability of a wrong classification of a new sample. It is based on an induction method which minimizes the upper limit of the generalization error related to uniform convergence, dividing the problem space using hyperplanes or surfaces, splitting the training samples into positive and negative groups and selecting the surface which keeps more samples.

K-Nearest Neighbours (KNN) [4]: This embodies one of the most simple learning methods. The training set is seen as composed of n -dimensional vectors and each element represents an n -dimensional space point. The classifier estimates the k nearest neighbours in the whole dataset based on an appropriate distance metric (Euclidian distance in the simplest case). The classifier checks the class labels of each selected neighbour and chooses the class that appears most in the label set.

Decision Trees (DT) [17]: This classifier uses a generalized "divide and conquer" strategy, splitting a complex problem into a succession of smaller sub-problems, and forming a hierarchy of connected internal and external nodes. An internal node is a decision point determining, according to a logical test, the next node reached. If this is an external node, the test sample is assigned to the class associated with that node.

Optimized IREP (Incremental Reduced Error Pruning) (JRip) [10]: The Decision Tree usually uses pruning techniques to decrease the error rates of a dataset with noise, one approach to which is the Reduced Error Pruning method. Specifically, we use Incremental Reduced Error Pruning (IREP). The IREP tries to divide to conquer. This algorithm uses a set of rules which, one

by one, are tested to check whether a rule matches, all samples related to that rule then being deleted. This process is repeated until there are no more samples or the algorithm returns an unacceptable error. Our algorithm uses a delayed pruning approach to avoid unnecessary pruning, resulting in a JRip procedure.

Naive Bayesian Learning (NBL) [9]: This algorithm relates to a simple probabilistic classifier based on the application of Bayes theorem with the assumption of strong independence. The principle is to estimate the conditional probability of each class label with respect to the test sample. In this method, it is assumed that each attribute is independent of the others.

3 Experimental Study

In order to determine the performance of the classifiers described, two databases of biometric samples were chosen, containing respectively, samples of hand-written signatures and fingerprint images. Section 3.1 describes the signature database and the results of an empirical investigation of classification of this data, while Section 3.2 describes a similar investigation with respect to the fingerprint samples.

3.1 Signature Database

The database contained signature samples collected as part of a BTG/University of Kent study [11] from 359 volunteers (129 male, 230 female) from a cross-section of the general public. The capture environment was a typical retail outlet, providing a real-world scenario in which to acquire credible data. There are 7428 signature samples in total, where the number of samples from each individual varies between 2 and 79, according to the distribution shown in Table 1.

Gender	2-10 samples	11-30 samples	31-50 samples	51-79 samples
Female	54	148	23	5
Male	42	66	22	9

Table 1. Distribution of sample set sizes

The data was collected using an A4-sized graphics tablet with a density of 500 lines per inch. For our study 18 representative features were extracted from each sample. These features were:

- Execution Time: The time required to execute the signature.
- Pen Lift: The number of times the pen was removed from the tablet during the execution process.
- Signature Width: The width of the image in mm.
- Signature Height: The height of the image in mm.

- Height to Width Ratio: The division of the signature height by the signature width.
- Average Horizontal Pen Velocity in X: The pen velocity in the x plane across the surface of the tablet.
- Average Horizontal Pen Velocity in Y: The pen velocity in the y plane.
- Vertical Midpoint Pen Crossings: The number of times the pen passes through the centre of the signature.
- M_{00} : Number of points comprising the image.
- M_{10} : Sum of horizontal coordinate values.
- M_{01} : Sum of vertical coordinate values.
- M_{20} : Horizontal centralness.
- M_{02} : Vertical centralness.
- M_{11} : Diagonality - indication of the quadrant with respect to centroid where image has greatest mass.
- M_{12} : Horizontal Divergence - indication of the relative extent of the left of the image compared to the right.
- M_{21} : Vertical Divergence - indication of the relative extent of the bottom of the image compared to the top.
- M_{30} : Horizontal imbalance - location of the centre of gravity of the image with respect to half horizontal extent.
- M_{03} : Vertical imbalance - location of the centre of gravity of the image with respect to half vertical extent.

Because of the nature of the data collection exercise itself, the number of samples collected differs considerably across participants. We impose a lower limit of 10 samples per person for inclusion in our experimentation, this constraint resulting in a population of 273 signers and 6956 signatures for experimentation. Table 2 shows the performance of the best individual classifiers with respect to the signature database, where the classifier configurations used were chosen taking into account the smallest mean overall error rate. As can be seen, the error delivered by the FuzzyMLP classifier is the smallest of the algorithms tested, although a very wide variation in achievable performance is observed. Arranging performance indices in decreasing order also reveals a general relationship between error rate performance and classifier complexity.

Table 3 presents a more detailed analysis of the performance results, recording separately the false positive and false negative error rates, and sub-dividing the test population into four different broad age groups. This shows that, in general, the false negative error rate exceeds the false positive rate. However, it is especially interesting to note (the sometimes quite marked) performance differences between the different age groups, especially if the youngest and oldest groupings are compared.

These results are very interesting, both because they again reveal significant diversity in relation to the performance characteristics of different classifier approaches, but also because they point to a changing performance profile when considered on an age-related basis. We observe error rates rising in the elderly population group as compared with the younger signers, a factor which is apparent both for false positive and false negative errors, although the increase is

Method	Error Mean \pm Standard Deviation
FMLP	8.47 \pm 2.92
MLP	9.88 \pm 2.81
RBF	12.51 \pm 2.97
SVM	12.78 \pm 4.21
JRip	15.72 \pm 3.12
NBL	18.74 \pm 2.45
DT	17.27 \pm 3.52
KNN	20.71 \pm 3.18

Table 2. Error Mean \pm Standard Deviation of the Signature Database

Method	18-25y		26-40y		41-60y		over 60y	
	fp	fn	fp	fn	fp	fn	fp	fn
FMLP	0.51	1.79	0.27	1.55	0.28	1.11	0.99	1.97
MLP	0.73	1.48	0.41	1.07	0.53	1.09	1.76	2.81
RBF	0.93	2.11	0.45	1.69	0.85	1.43	2.07	2.98
SVM	0.92	2.81	0.51	1.60	0.37	1.94	1.84	2.79
JRip	0.97	3.69	0.34	2.18	0.41	2.48	1.17	4.48
NBL	1.83	3.94	0.87	2.12	0.92	2.51	2.86	5.07
DT	1.67	2.85	1.02	1.59	0.83	2.25	2.78	4.28
KNN	2.91	3.85	1.57	2.16	1.14	2.27	2.28	4.53

Table 3. False Positive (fp) and False Negative (fn) of the Signature Database

generally more marked in the former case. It is also seen that the less powerful classification algorithms smooth out these age-related differences, although against a background of generally poorer error rate performance.

3.2 Fingerprint Database

The database used for our study of fingerprint data was that compiled for the Fingerprint Verification Competition 2002 [14]. This in fact comprises four different (sub)-databases (designated DB1, DB2, DB3 and DB4), three of them containing images of "live" prints acquired with different sensors, and the fourth containing synthetically generated fingerprint images.

	Sensor Type	Image Size	Resolution
DB1	Optical (TouchView II - Identix)	388x374 (142 Kpixels)	500 dpi
DB2	Optical (FX2000 - Biometrika)	296x560 (162 Kpixels)	569 dpi
DB3	Capacitive (100 SC - Precise Biometrics)	300x300 (88 Kpixels)	500 dpi
DB4	Synthetic (SFInGe v2.51)	288x384 (108 Kpixels)	about 500 dpi

Table 4. Devices used in the Fingerprint acquisition

The evaluation of the real datasets was performed in three groups of 30 people each. There were three sessions where prints from four fingers per person were collected, and the images included variations in the collection conditions, such as varying types of distortion, rotation, dry and moist fingers. For each dataset, a subset of 110 separate fingers, with eight impressions per finger, was included (880 samples at all). Each dataset is divided in two sets, set A (800 samples) and set B (80 samples). The individuals donating the prints are different in each dataset. Table 4 records the sensor technologies and other relevant information for each database.

Method	DB1	DB2	DB3	DB4
FMLP	16.09 \pm 3.61	9.46 \pm 2.94	13.71 \pm 3.61	9.90 \pm 2.59
MLP	20.66 \pm 3.64	10.02 \pm 2.25	16.94 \pm 3.29	10.98 \pm 3.59
RBF	17.78 \pm 3.48	10.19 \pm 3.64	16.09 \pm 4.53	14.8 \pm 2.67
SVM	24.94 \pm 4.89	17.03 \pm 2.81	21.97 \pm 6.00	17.69 \pm 3.67
JRip	23.02 \pm 5.47	15.79 \pm 3.91	13.81 \pm 4.67	16.89 \pm 3.99
NBL	21.27 \pm 3.71	16.21 \pm 2.77	14.83 \pm 3.16	17.44 \pm 2.99
DT	21.36 \pm 4.61	16.00 \pm 3.67	14.34 \pm 5.02	17.69 \pm 3.69
KNN	30.16 \pm 6.59	23.12 \pm 2.78	26.74 \pm 5.88	23.79 \pm 2.87

Table 5. Error Mean \pm Standard Deviation of the Fingerprint Database

The minutiae were extracted using the NFIS2 (NIST Fingerprint Image Software 2) [1]. Each minutia is represented by eight indicators, as follows:

- Minutia Identifier
- X-pixel Coordinate
- Y-pixel Coordinate
- Direction
- Reliability Measure
- Minutia Type
- Feature Type
- Integer Identifier of the feature type

As each finger presents a different number of detectable minutiae, while the classifiers adopted need a common number of entries, it is necessary to fix the number of minutia. During the construction of the dataset, where a sample contains fewer minutiae than the chosen number, random non-real data was added to compensate. On the other hand, where a sample contains too great a number of minutiae, the excess minutiae were randomly discarded.

Table 5 shows the error rates obtained with the fingerprint data (cf. Table 2). As was the case with the signature-based experiment, the mean error delivered by the FuzzyMLP classifier is smaller than all other classifiers, but in this case the pattern of classification performance across the whole tested range differs from the previous experiment. We note, however, that the KNN classifier again

	DB1		DB2		DB3		DB4	
Method	fp	fn	fp	fn	fp	fn	fp	fn
FMLP	4.18	11.91	2.97	6.49	2.72	10.99	1.86	8.04
MLP	2.73	17.93	3.55	6.47	4.55	12.39	1.21	9.77
RBF	3.86	13.92	3.94	6.25	1.21	14.88	5.25	9.55
SVM	6.07	18.87	3.77	13.26	2.30	19.67	3.97	13.72
JRip	7.03	15.99	6.13	9.66	1.89	11.92	4.30	12.59
NBL	2.63	18.64	5.44	10.77	4.20	10.63	4.96	12.48
DT	2.93	18.43	6.29	9.71	3.60	10.74	2.76	14.93
KNN	8.46	21.7	7.13	15.99	5.02	21.72	6.72	17.07

Table 6. False Positive (fp) and False Negative (fn) of the Fingerprint Database

performs the poorest. This behaviour demonstrates that this data is somewhat more challenging than the signature case, largely because of the problem of missing minutiae in the samples, but also reveals common trends in classifier performance across modalities.

Table 6 shows error rates broken down into false positive and false negative rates. The false positive rate is greater than the false negative, and performing the t-test between the two classifiers with the smaller error means gives the figures shown in Table 7. This shows that the FuzzyMLP is statistically more accurate than the classifiers returning the second largest correct mean.

Database	Classifiers Tested	<i>p</i> Value
DB1	FMLP x RBF	0.000451
DB2	FMLP x MLP	0.066
DB3	FMLP x JRip	0.433
DB4	FMLP x MLP	0.00779

Table 7. T-test to Fingerprint Database

The available literature reports a number of studies [2] [3] [7] [8] using this database, with a particular focus on DB3 because of its particularly poor image quality. Our study shows some particularly interesting characteristics in relation to these studies, enhancing current insights into this important classification task domain.

4 Discussion and Conclusions

In this paper we have reported on an empirical study of classifier performance in typical biometric data classification tasks. Although some caution needs to be exercised in interpreting such results, especially in generalizing specific indicators, this study provides some pointers to useful practical conclusions, as follows:

- We have provided some empirical data which demonstrates the wide variability in identification performance in relation to classifier selection for a given modality. This is seen to be the case both when the principal index of performance is absolute overall error rate and, perhaps most significantly, also when the balance between False Acceptance and False Rejection is considered.
- Although caution is advisable when pointing to any individual classifier as representing a "best" choice, our experiments do reveal some general trends concerning the relative merits of different classification approaches which, while not absolute, may be useful pointers to selection strategies.
- A finer-grained analysis of performance within a specific modality can also generate useful practical insights into the relation between lower-level factors and performance returned using different classification approaches. In relation to the signature modality, for example, even our basic analysis of different age profiles within a population reveals important information about changing patterns of vulnerability with respect to system performance indicators across the age spectrum. This could be very significant in system optimisation in a number of application scenarios.
- Multiclassifier solutions to single modality configurations are under-represented in the literature, and yet the multiclassifier methodology is widespread and often very effective in many application domains. Our empirical study provides relevant information to inform further investigation of this approach to enhancing identification performance.
- Despite the fact that multiclassifier systems can combine the benefits of many classifiers, they do not necessarily provide entirely "intelligent" solutions. It may be advantageous for the classifiers to be more interactive taking account of their individual strengths and weaknesses. Multiagent systems offer such a possibility, and our results provide a starting point for designing a novel solution based on such an operating principle.
- Multibiometric solutions are now widely recognised to offer advantages not only in enhancing overall system performance, but also, significantly, in offering greater flexibility and user choice in system configuration. This study provides some initial insights into how to match classifiers and modality-specific data in determining an optimal configuration. Moreover, although there is now an extensive literature on modality combination, adopting the signature as one of the target modalities is a relatively little used option, and our benchmark performance characterisation can provide a starting point for a productive study of optimal modality selection.

This study therefore both provides some quantitative data to characterise some common approaches to classifier implementation for application to practical scenarios in biometrics, and sets out some possibilities for developing more sophisticated and effective strategies for developing enhanced practical systems in the future.

Acknowledgment

The authors gratefully acknowledge the financial support given to Mrs Abreu from CAPES (Brazilian Funding Agency) under grant BEX 4903-06-4.

References

1. Nist Fingerprint Image 2. User's guide to.
2. M. M. A. Allah. Artificial neural networks based fingerprint authentication with clusters algorithm. *Informatica (Slovenia)*, 29(3):303–308, 2005.
3. M. M. A. Allah. A novel line pattern algorithm for embedded fingerprint authentication system. *ICGST International Journal on Graphics, Vision and Image Processing*, 05:29–35, March 2005.
4. S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6):891–923, 1998.
5. M. D. Buhmann. *Radial Basis Functions*. Cambridge University Press, New York, NY, USA, 2003.
6. A. M. P. Canuto. *Combining Neural Networks and Fuzzy Logic for Applications in Character Recognition*. PhD thesis, Department of Electronics, University of Kent, Canterbury, UK, Maio 2001.
7. Y. Chen, S. C. Dass, and A. K. Jain. Fingerprint quality indices for predicting authentication performance. In *AVBPA*, pages 160–170, 2005.
8. S. Chikkerur, A. N. Cartwright, and V. Govindaraju. Fingerprint enhancement using stft analysis. *Pattern Recognition Letter*, 40(1):198–211, 2007.
9. C. Elkan. Boosting and naive bayesian learning. Technical report, 1997.
10. J. Fürnkranz and G. Widmer. Incremental reduced error pruning. In *ICML*, pages 70–77, 1994.
11. R. M. Guest. The repeatability of signatures. In *IWFHR '04: Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04)*, pages 492–497, Washington, DC, USA, 2004. IEEE Computer Society.
12. Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998.
13. Friedrich Leisch, Lakhmi C. Jain, and Kurt Hornik. Cross-validation with active pattern selection for neural-network classifiers. *IEEE Transactions on Neural Networks*, 9(1):35–41, 1998.
14. D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. Fvc2002: Second fingerprint verification competition. In *ICPR '02: Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, volume 3, page 30811, Washington, DC, USA, 2002. IEEE Computer Society.
15. T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
16. C. Nello and S.-T. John. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000.
17. J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
18. F. Rosenblatt. The perception: a probabilistic model for information storage and organization in the brain. pages 89–114, 1988.

Promoting diversity in Gaussian mixture ensembles: an application to signature verification

Jonas Richiardi¹, Andrzej Drygajlo¹, and Laetitia Todesco¹

Institute of Electrical Engineering
Swiss Federal Institute of Technology Lausanne
Switzerland
{jonas.richiardi,andrzej.drygajlo}@epfl.ch,
<http://scgwww.epfl.ch/>

Abstract. Classifiers based on Gaussian mixture models are good performers in many pattern recognition tasks. Unlike decision trees, they can be described as stable classifier: a small change in the sampling of the training set will produce not a large change in the parameters of the trained classifier. Given that ensembling techniques often rely on instability of the base classifiers to produce diverse ensembles, thereby reaching better performance than individual classifiers, how can we form ensembles of Gaussian mixture models? This paper proposes methods to optimise coverage in ensembles of Gaussian mixture classifiers by promoting diversity amongst these stable base classifiers. We show that changes in the signal processing chain and modelling parameters can lead to significant complementarity between classifiers, even if trained on the same source signal. We illustrate the approach by applying it to a signature verification problem, and show that very good results are obtained, as verified in the large-scale international evaluation campaign BMEC 2007.

1 Introduction

Successful ensembling methods such as bagging [3] and boosting [5] rely on the fact that the ensemble member classifiers are *unstable*, that is, a small change in the sampling of the training set will produce a large change in the trained classifier. Unstable classifiers include decision trees and neural networks [3], while others such as naïve Bayes are considered stable [8]. In reality, there is a continuum of stability, in the sense that the *amount* of output change incurred by different classifiers with respect to changes in the training set is not simply binary (“stable” or “unstable”) [2].

Training several unstable classifiers with different sampling of the training set is one way to produce an ensemble that is diverse. The hope is that the training procedure produces classifiers whose output is complementary: they yield erroneous outputs for different samples. By combining these classifiers, the total variance can be reduced, typically leading to reductions in expected error rates.

In many applications dealing with real-life signals, a classifier that systematically yields good results is the Gaussian mixture model (see e.g. [13]). Example applications are speaker verification [16] or face recognition [21]. Leaving out effects of critically small training sample sizes with respect to the model complexity, Gaussian mixture models can be considered as stable classifiers. Given that multiple-classifier systems can outperform single-classifier systems on a large number of tasks and datasets [12], it would seem beneficial to build ensembles of Gaussian mixture classifiers. However, as pointed out above, diversity is an important factor for successful ensembling. How, then, can we increase diversity in ensembles of stable classifiers?

Recent work has shown that adding components to stable classifiers before ensembling could improve results over standard techniques such as bagging for these classifiers. For example, in the Random Oracle technique applied to naïve Bayes classifiers [19], the training set is split at random between the two classifiers, and at test time the base classifier is also selected at random. Another technique based on a hybrid of naïve Bayes and decision trees, called Levelled Naïve Bayesian Trees [22], is to grow a decision tree whose leaves are naïve Bayes classifiers. The hope there is that the naïve Bayes classifiers will inherit the instability of the decision tree growing procedure, and make them more amenable to boosting.

In this paper, to optimise the coverage of the ensemble, we propose instead to act at different levels of the pattern recognition processing chain of individual classifiers in order to increase diversity in ensembles of Gaussian mixture classifiers, and note that this does not prevent the application of other destabilising techniques. We should also note that, while it seems “diversity” is a desirable property of classifier ensembles in order to reduce error rate, there is no consensus on how to measure it and how it relates to ensemble performance [11], although theoretical work in this area is progressing [14].

The rest of this paper is laid out as follows: In Section 2 we present in more details techniques that can be used to increase diversity in ensembles of stable classifier. In Section 3, we show the detailed application of these principles to a multiple-classifier signature verification system based on Gaussian mixture models. In Section 4 we provide experimental results on a signature verification database, and Section 5 concludes the article.

2 Increasing diversity in ensembles of stable classifiers

A pattern recognition systems consists of a front-end responsible for extracting features, a training procedure to learn the parameters of the classifier, and a testing algorithm to obtain soft or hard output from the classifier. We will now examine these levels in more details and how they can be modified to influence the output of a classifier, which in turn can promote diversity in an ensemble. In the application field of biometrics, some of these techniques fall under the general heading of “multibiometrics” [20].

2.1 Changes to the front-end

The front-end to pattern recognition systems uses a signal processing chain that starts with real-world analogue signals. A schematic view is shown on Figure 1.



Fig. 1. Front-end for pattern recognition.

Changes in any of the processing steps will affect all other steps further downstream, and lead to various amounts of classifier diversity. Even within the same modality (say, infrared images), changing the sensor at the **signal acquisition** stage can lead to significant differences between classifiers. In this regard, multimodal pattern recognition can be seen as a way to obtain diverse ensembles.

The **pre-processing** performed on the data can have a large influence on the feature extraction process. Filtering, denoising, imputing missing data and other linear and non-linear transformations of the digitised signal can lead to significant differences further down the processing chain.

The representation of the signal as vectors of features typically involves a non-linear transformation of the pre-processed signal. For example, the use of Fourier transforms and related transforms such as the DCT at the **feature extraction** stage change signal representation and may permit the extraction of features that lead to classifiers complementary to those trained on other signal representations. This technique is used in many applications such as language recognition, where different parameterisations of speech are often combined [15], or fingerprint recognition, where minutiae can be combined with skin pores [10]. Even within the same signal representation, it is possible to use random feature subspace methods [7] to purposefully obtain diverse classifiers.

Finally, the **post-processing** stage, which typically consists of some form of statistical normalisation of the feature vectors (mean removal being typical in speech applications [6]), can also introduce important changes to the trained parameters of the classifier by applying linear or non-linear transformations to the original feature space.

2.2 Changes in the sampling of the training set

By our definition of stability, varying the sampling of the training set, a common strategy for achieving diversity in ensembles, will not be effective for increasing diversity in ensemble of stable classifiers (although see [19] for a more sophisticated approach). Thus, we propose to concentrate efforts on other parts of the pattern recognition system.

2.3 Change in model complexity

Classifiers implemented as statistical models (Gaussian mixture models, generative Bayesian networks) form a family in which the number of parameters has a great influence on classification results. For example, modifying the covariance matrix structure (say, from diagonal to full) can substantially alter the output of the classifier. Likewise, by modifying the number of hidden variables in a Bayesian network corresponding to the number of components in a mixture of Gaussians, and thereby changing the number of parameters in the model, it is possible to decorrelate stable models that are trained from feature vectors where everything else in the front-end (acquisition, pre-processing, feature extraction, post-processing, samplig of the training set) is identical.

2.4 Change in scoring procedure

The same model can be used to compute a score in different ways. Depending on the model type, this is a way to promote diversity. In this regard, the recent technique presented in [23], whereby a hidden Markov model is used to produce likelihood output and a Viterbi-related output which are then combined, can be seen as a way to exploit complementarity in classifier output. However, for GMMs, it is likely that gains obtained from combining all-components scoring with top-components-scoring¹ would be small.

3 Application: a Gaussian mixture ensemble for signature verification

In this section, we present an application of the techniques exposed in Section 2 to the problem of signature verification, where the Gaussian mixture model is one of the best-performing classifiers [17]. The goal is to train a diverse set of signature verification classifiers, so that they can be effectively combined. The Gaussian mixture ensemble we present consists of $L=6$ different Gaussian mixture model classifiers. In fact, since biometric verification problem can be cast as a series of 2-class problems, each of the U users is modelled by one of the U Gaussian mixture ensembles.

We do not use a measure of diversity based on the label (hard, binary) outputs of the classifiers [11], but rather the normalised mutual information between the scores (soft, continuous) outputs of the classifiers. We assume that having lower mutual information between pairs of classifiers is equivalent to having a higher diversity in the ensemble². We use the following definition for normalised mutual information:

¹ This is a common technique in speaker recognition [1], where high model orders and large datasets warrant the summing of *some* of the Gaussian components in the likelihood computation

² Using conditional mutual information would allow us to take into account the effect of already having included certain classifiers in the ensemble.

$$\bar{I}(S_{c_1}; S_{c_2}) \triangleq \frac{I(S_{c_1}; S_{c_2})}{\sqrt{H(S_{c_1})H(S_{c_2})}}, \quad (1)$$

where $I(S_{c_1}; S_{c_2})$ is the mutual information between the scores output of two classifiers, and $H(S_{c_l})$ is the entropy of the scores output of the l th classifier.

3.1 Preprocessing

On some low-power signature acquisition platforms such as personal digital assistants, data acquisition may produce missing values intermittently. Missing data is also a frequent occurrence in slow and fast strokes. In this case, an effective approach is to interpolate the missing data. By using different interpolation algorithms, or none at all, it is possible to introduce variability in the signal which will be reflected further down the processing chain. Figure 2 shows the result of two different interpolation methods on the same data. Looking at a single classifier, it is not obvious which interpolation method is the best in terms of accuracy.

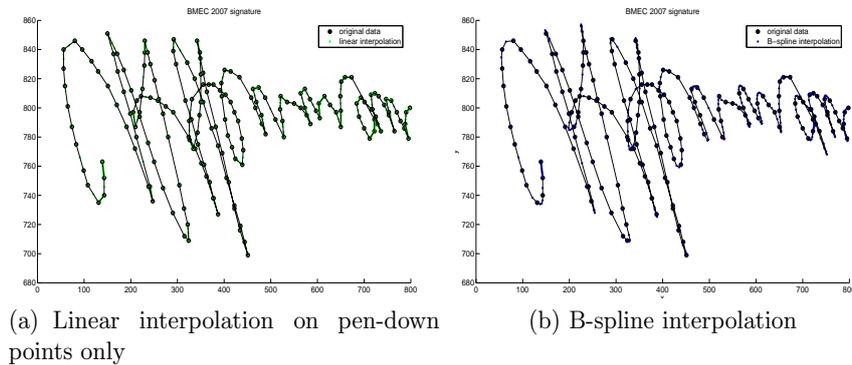


Fig. 2. Signature preprocessing for recovery of missing data on BMEC 2007

A second pre-processing technique that could lead to diversity is rotation normalisation. Indeed, in some situations, such as handheld device-based acquisition, it is likely that the orientation of the signature with respect to the horizontal axis of the acquisition surface can be very variable. We estimate the principal axis of the signature by eigendecomposition: The eigenvector associated with the largest eigenvalue indicates the axis of greatest variance. Again, from looking at the accuracy of a single classifier it is not obvious whether this really is of help, but it can be used to force diversity in an ensemble.

The preprocessing used by the local and global classifiers in our ensemble is detailed in Table 1.

3.2 Feature Extraction

In the parametric paradigm, local, segmental, or global parameters are computed from the pre-processed signals and used as features.

Local features are extracted at the same rate as the incoming signal: that is, each input sample data vector corresponds to a local feature vector.

Segmental features are extracted once the signature has been cut into segments. A segment typically consists of a sequence of points for which some definition of coherence holds.

Global features summarise some property of the complete observed signature; for instance the total signing time, pen-up to pen-down ratio, etc.

Changing the signal representation and combining the resulting classifiers is a common technique in pattern recognition, and has been applied also to signature verification [4]. Our Gaussian mixture ensemble consists of 5 classifiers trained on local features, and 1 classifier trained on global features (see Table 1).

3.3 Modelling

Diversity can be enforced in ensembles of Gaussian mixture models by changing the number of parameters used for the constituent classifiers, for instance by changing the type of covariance matrix (diagonal, full, spherical...), or by choosing a different number of Gaussian components in the mixture. A further way of increasing diversity is by using a MAP adaptation scheme instead of direct training.

3.4 Diversity in the ensemble

The 5 GMM classifiers based on local features, denoted $GL_{1...5}$, and the GMM classifier based on global features, denoted GG , use the specific combinations of preprocessing, feature extraction, and model orders shown in Table 1. In the table, LI refers to linear interpolation, while B-S refers to B-spline interpolation. *Rotation* indicates whether rotation normalisation is performed or not. The feature sets are as follows: feature set 1 comprises $\{x_t, y_t, \Delta, \Delta\Delta\}$, where x_t and y_t are the sampled horizontal and vertical position of the pen. The Δ and $\Delta\Delta$ features are numerically approximated first, respectively second derivatives of the base features. Feature set 2 is $\{x_t, y_t, \theta_t, v_t, \Delta, \Delta\Delta\}$, where θ_t is the writing angle and v_t is the instantaneous writing speed. Feature set 3 is $\{x_t, y_t, z_t, \Delta, \Delta\Delta\}$, where z_t is a binary variable representing pressure. Feature set 4 comprises 11 global features, described in [18]. Lastly, different number of components are used in the mixture, denoted *user model order*.

In terms of classifier output, these changes result in a diverse ensemble of GMMs, with complementarity clearly showing on Figure 3. As could be expected, the different parameterisation of the signal (local or global) result in the largest diversity, but it can also be observed that changing the model order or the preprocessing can also lead to important changes in classifier output. To put the results in perspective, the normalised mutual information between a vector \mathbf{x}

Name	GL_1	GL_2	GL_3	GL_4	GL_5	GG
Interpolation	LI	B-S	LI	LI	B-S	LI
Rotation	y	n	y	n	y	n
feature set	1	1	1	2	3	4
user model order	24	36			2	
world model order	4					

Table 1. Details of classifier in the ensemble.

consisting of 1000 samples drawn at random from a uniform distribution between 0 and 1 and the vector-valued $\sin(\mathbf{x})$, a near-linear relationship, is 0.75. The normalised mutual information between two vectors of dimension 1000 randomly drawn from a uniform distribution between 0 and 1 is 0.02. Thus, it can be seen that significant reductions in dependence between classifiers can be achieved by applying the approach proposed here: for example classifiers GL_1 and GL_3 have a normalised mutual information of 0.41, while the only difference between them is the model order (and the random initialisation of the EM algorithm).

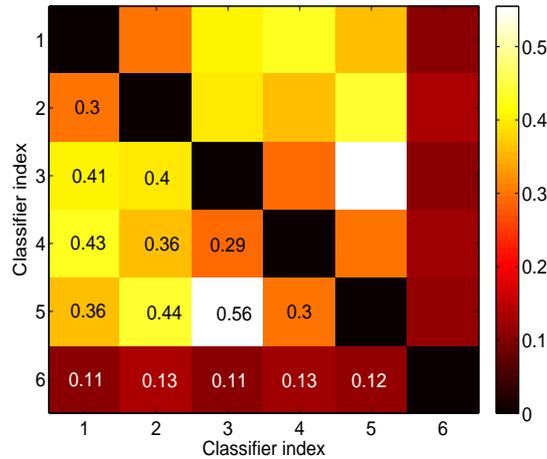


Fig. 3. Mutual information between classifiers in the ensemble. Note that the diagonal (equivalent to the entropy of each classifier) has been set to 0 for enhanced contrast.

4 Verification experiments and results

4.1 Database

The BMEC2007 development database contains 50 users, each with two sessions, and is part of the larger BioSecure DS3 dataset. For each user, the first session

contains 5 genuine signatures and 10 skilled forgeries³. The second session contains 15 genuine signatures and 10 skilled forgeries. Signatures are acquired on a low-power mobile platform (Ipaq PDA). This means that some data is missing, and interpolation approaches outlined in Section 3.1 have to be applied. Furthermore, the orientation of the signatures is haphazard. The acquisition platform only captures binary pressure (on/off) and x,y signals. No pen orientation information is available. The low quality of the data explains why error rates are in general high on this database compared to other signature databases.

4.2 Protocol

For each user, We train their set of classifiers ($GL_{1...5}$ and GG) on the 5 genuine signatures of the first session. We then run these classifiers on the remaining held-out test data. Thus, for each user we obtain 15 genuine and 20 skilled forgery scores, resulting in a total of 750 genuine signature scores and 1000 skilled forgery scores.

The ensemble classifier (in the present case a simple mean rule, but similar results are obtained using logistic regression) is then trained and tested with this score data using 5-fold cross-validation.

4.3 Results

Glancing at Figure 4, it appears that the local classifiers in the ensemble offer approximately the same performance, while the global classifier trails behind. By ensembling local classifiers via the mean rule, it is already possible to substantially lower the error rate, indicating that our coverage optimisation approach based on changes in preprocessing, feature subsets, and modelling complexity is effective. Further adding a global classifier, itself with different features and modelling complexity, yields improved performance. This could be expected given that global information is complementary with local information, and that time information (signature length) is incorporated in the global feature set. While not reported here, we have performed experiments on other signature databases with similar results. It is interesting to note that, while classifiers GL_3 and GL_4 have virtually identical performances, their mutual information is low (0.3); this is to be accounted for mainly by the rotation normalisation and the inclusion of tangent angles in one feature set. None of them stands out in isolation, but they can be usefully combined because of their diversity. It is certainly possible to reduce the complexity of this ensemble by removing a few local classifiers, while still preserving an adequate accuracy.

This ensemble performed well in the BMEC 2007 competition, comprising a database 430 users, and has taken first place for random forgeries (about 4.0% EER), second place for skilled forgeries (about 13.6% EER), and first place for synthetic forgeries (about 10.7% EER).

³ These forgeries fall between levels 6 and 8 in [9, Table 3], as the forger has no visual contact with the victim, but is allowed to see several times the dynamics of signing.

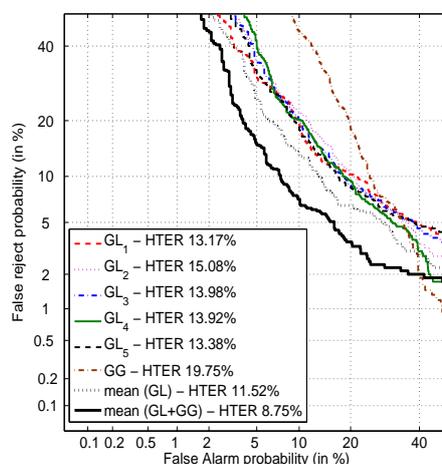


Fig. 4. Verification results (skilled forgeries) for base classifiers ($GL_{1..5}$ and GG) and Gaussian mixture ensemble with only local classifiers (mean GG) and local and global classifiers (mean $GL + GG$).

5 Conclusions

In biometric verification applications, Gaussian mixture models are generally top performers. Other classifiers commonly used in pattern recognition, such as decision trees or random forests, are not often used as base classifiers. We have shown that despite their being categorised as stable, Gaussian mixture models can serve as base classifiers in ensembles if coverage is optimised adequately. To this end, the signal processing chain and other components of the pattern recognition pipeline has to be modified to introduce variability. While the resulting classifiers have roughly the same accuracy, they are complementary and can be usefully combined in an ensemble.

References

1. Jean-François Bonastre, Frédéric Wils, and Sylvain Meignier. ALIZE, a free toolkit for speaker recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 737–740, Philadelphia, USA, March 2005.
2. Olivier Bousquet and André Elisseeff. Stability and generalization. *J. of Machine Learning Research*, 2:499–526, 2002.
3. Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, August 1996.
4. J. Fierrez-Aguilar, L. Nanni, J. Lopez-Peñalba, J. Ortega-Garcia, and D. Maltoni. An on-line signature verification system based on fusion of local and global information. In *Proc. 5th IAPR Int. Conf. on Audio and Video-based Biometric Person Authentication (AVBPA)*, pages 523–532, 2005.
5. Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, September 1995.

6. S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 29(2):254–272, April 1981.
7. T.K. Ho. The random space method for constructing decision forests. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.
8. Zoë Hoare. Landscapes of naïve Bayes classifiers. *Pattern Analysis and Applications*, 11(1):59–72, 2007.
9. ISO/IEC JTC 1/SC 37 Biometrics. TR 19795-3, biometric performance testing and reporting, part 3: Modality specific testing. Technical report, International Standards Organization, 2007.
10. K. Kryszczuk, P. Morier, and A. Drygajlo. Study of the distinctiveness of level 2 and level 3 features in fragmentary fingerprint comparison. In *International Conference on Computer Vision, Biometric Authentication Workshop*, Prague, Czech Republic, May 2004.
11. Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, May 2003.
12. Ludmila Ilieva Kuncheva. *Combining Pattern Classifiers*. Wiley and sons, 2004.
13. Geoffrey J. McLachlan and Kaye E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, 1987.
14. Julien Meynet and Jean-Philippe Thiran. Information theoretic combination of classifiers with application to adaboost. In *Proc. 7th Int. Workshop on Multiple Classifier Systems*, pages 171–179, 2007.
15. Christian Müller and Joan-Isaac Biel. The ICSI 2007 language recognition system. In *Proc. Odyssey 2008: The Speaker and Language Recognition Workshop*, Stellenbosch, South Africa, January 2008.
16. D.A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.
17. J. Richiardi and A. Drygajlo. Gaussian mixture models for on-line signature verification. In *Proc. ACM SIGMM Multimedia, Workshop on Biometrics methods and applications (WBMA)*, pages 115–122, Berkeley, USA, Nov. 2003.
18. Jonas Richiardi, Hamed Ketabdardar, and Andrzej Drygajlo. Local and global feature selection for on-line signature verification. In *Proc. IAPR 8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, volume 2, pages 625–629, Seoul, Korea, August-September 2005.
19. Juan Rodríguez and Ludmila Kuncheva. Naïve bayes ensembles with a random oracle. In *Proc. 7th Int. Workshop on Multiple Classifier Systems (MCS)*, pages 450–458, 2007.
20. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Springer, 2006.
21. Conrad Sanderson. *Automatic Person Verification Using Speech and Face Information*. PhD thesis, Griffith University, Queensland, Australia, 2002.
22. Kai Ting and Zijian Zheng. Improving the performance of boosting for naive bayesian classification. In *Proc. 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 296–305, 1999.
23. Bao Ly Van, Bao Ly Van, S. Garcia-Salicetti, and B. Dorizzi. On using the viterbi path along with hmm likelihood information for online signature verification. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 37(5):1237–1247, 2007.

Advanced Studies on Reproducibility of Biometric Hashes

Tobias Scheidat¹, Claus Vielhauer^{1,2}, Jana Dittmann¹

¹ Otto-von-Guericke University of Magdeburg, Universitätsplatz 2,
39106 Magdeburg, Germany

² Brandenburg University of Applied Sciences, PSF 2132,
14737 Brandenburg, Germany

¹ {tobias.scheidat | claus.vielhauer | jana.dittmann}@iti.cs.uni-magdeburg.de
² claus.vielhauer@fh-brandenburg.de

Abstract. The determination of hashes based on biometric data is a recent topic in biometrics as it allows to handle biometric templates in a privacy manner. Applications could be the generation of biometric templates for authentication or of cryptographic keys based on biometric traits. Depending on the application, there are different requirements with regard to possible errors. On one side, authentication performance based on biometric hashes as feature representation can be measured by common biometric error rates such as FNMR, FMR and EER. Thus, generated hashes for each single person have to be only similar in a certain degree, but not necessarily identical. On the other side, biometric hashes for cryptographic issues have to be identical and unique for each individual, although measured data from same person differs from one acquisition to next, or data from different people may be similar. Therefore, we suggest three measures to estimate the reproducibility performance of biometric hash algorithms for cryptographic applications. To prove the concept of the measures, we provide an experimental evaluation of an online handwriting based hash generation algorithm using a database of 84 users and different evaluation scenarios.

Keywords: Biometrics, biometric hashing, collision, handwriting, measures, reproducibility, semantic fusion, verification

1 INTRODUCTION

In current biometric research, the generation of hash values based on biometric input is a recent topic. One goal of biometric hashing is the determination of a stable hash value based on a biometric trait of one person from its fuzzy input data in order to assure either authenticity and integrity, or confidentiality and privacy of biometric information. Another aim can be the generation of unique individual values for cryptographic purposes ([1]), since the biometric information of a person is available

anytime and anywhere, without the need to remember secret information or to present a special token.

In the following, a small selection from the variety of publications related to biometric hashing is presented, without neglecting others. In [2] the authors present a method to calculate a cryptographic key based on a spoken password. Therefore, a 12-dimensional vector of cepstral coefficients is used as well as an acoustics model, which is speaker dependent. Based on these components, segmentation is carried out in order to create different types of features as basis of a so called feature descriptor which can be used as hash value. The biometric hashing method described by Vielhauer et al. in [3] is based on online handwriting biometrics and determines a feature vector of statistical parameters. These parameters are transformed into a hash value space using an interval mapping function, which results in a hash vector as feature vector representation. This method is described in more detail in section 2, since it was used as reference algorithm for the evaluation in this paper. Further methods for biometric hash generation can be found also for other biometric modalities, e.g. for face [4], fingerprint [5] or DNA [6].

This paper is structured as follows: The next section discusses relations between cryptographic and biometric hash functions and introduces the Biometric Hash algorithm, which is used as reference algorithm for our experimental evaluation. In the third section, new measurements are described to estimate the reproducibility performance of a biometric hash function motivated from [7]. The fourth section explains a fusion strategy of combining biometric hashes based on different handwritten contents. The evaluation database, methodology and the results with regard to biometric error rates and hash reproducibility are described in the fifth section. The last section concludes this paper and gives an overview of future work in this field of biometric research.

2 BIOMETRIC HASHING

Since the idea of a biometric hashing function is based on the principles of cryptographic hashing, the first part of this section discusses differences and similarities of cryptographic and biometric hash functions. In the second part, the reference algorithm used in our experimental evaluation is reintroduced shortly.

2.1 Cryptographic hash functions vs. biometric hash functions

A cryptographic hash function ($h: A \rightarrow B$) has to fulfill different requirements ([8]): It has to be a so-called one-way function that provides the property of *irreversibility*, which describes the computational impossibility to determine any input data a from a hash value $h(a)$. Further, the *reproducibility* property of a hash function has to ensure that if any input data a and a' are equal, then also the output data $h(a)$ and $h(a')$ are equal. Contrariwise, in case a and a' are not equal, the corresponding hashes $h(a)$ and $h(a')$ have to be unequal. This requirement is called *collision resistance*. A fourth requirement of cryptographic hashes is the *bit sensitivity*. It states that small changes

in the input data a (e.g. by alternating one bit) should lead to a big change in the output data $h(a)$.

Biometric hash functions should be also one-way functions to avoid obtaining the private user-related or reliable biometric input data from hashes. However, since biometric data is varying each time of acquisition even for the same user and trait (intra-class variability), and data of different people may be similar (inter-class similarity), reproducibility and collision resistance have to be redefined for biometric hashing: On one side, *reproducibility* for the purpose of biometric hashing means the *identical hash reproduction for the same person and trait*, although the input data *varies within given bounds*. On the other side, the *collision resistance* of biometric hash functions describes the *ability to distinguish between (similar) data from different persons to generate different individual and unique hashes*. Consequently, due to the intra-class variability and inter-class similarity, the bit sensitivity property of cryptographic hashes cannot be mapped into the biometric hash methodology.

2.2 Biometric Hash algorithm for online handwriting biometrics

This subsection describes our Biometric Hash reference algorithm (see [3], [9]) based on online handwriting. Since we developed the new measures to quantify the degree of changes in an optimization process of the Biometric Hash algorithm, we use it as reference algorithm for our exemplarily evaluation based on these new measures. Figure 1 shows on the left side the enrollment process of the Biometric Hash algorithm. The first input data is a set of n raw data samples (D_1, \dots, D_n) derived from the handwriting acquisition sensor, e.g. tablet PC or PDA.

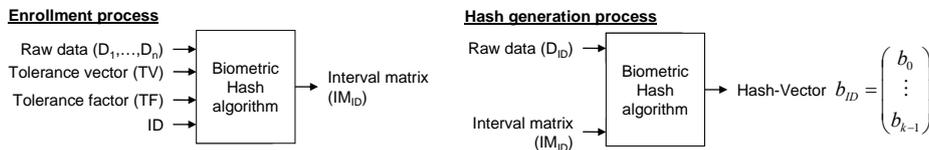


Fig. 1. Enrollment and hash generation processes of the Biometric Hash algorithm [ViSM2002]

The aim of the enrollment process is to generate a so-called interval matrix IM for each user based on its raw data and several parameters. Generally, each raw data sample D_i ($i=1, \dots, n$) consists of a temporarily dependent sequence of physical values supported by the device, such as pen tip coordinates $x(t)$ and $y(t)$, pressure $p(t)$ and pen orientation angles Altitude ($\Phi(t)$) and Azimuth ($\Theta(t)$). During the enrollment process, for each of the raw data samples D_i derived from a person, a statistical feature vector is determined with a dimensionality of k ($k=69$ in the current implementation). IM stores for each feature the length of an interval and an offset, where both values are calculated based on the intra-class-variability of the person, by using his/her statistical feature vectors. To parameterize the hash generation, the tolerance vector TV is used. The TV supports an element wise parameterization of the statistical features during the generation of hash values by the so-called interval mapping function. Thus, the dimensionality of TV is also k . The TV can be determined for each user individually or globally by a group of persons, either based on the registered users or a disjoint user set. The third input data is the tolerance factor TF as

global hash generation parameter, which is a scalar value. Using the TF , it is possible to scale the mapping intervals for all feature components globally by one factor, thus affecting both reproducibility and collision resistance, where increasing values of TF lead to the tendency of increasing reproducibility at cost of increasing collision probabilities. The user's identity ID is the fourth input for the enrollment process, which is linked to the reference data. Note that in our context, reference data is the output of the Biometric Hash algorithm's enrollment mode in form of the interval matrix IM_{ID} that provides information for the mapping of the individual statistical features to the corresponding hash values, but neither the original biometric input nor the actual feature vectors. The right side of Figure 1 shows the hash generation process of the Biometric Hash algorithm. Here, the input data consists of only one single raw data sample D_{ID} and the interval matrix IM_{ID} of a claimed identity ID . The raw data D_{ID} is used to determine a k -dimensional statistical feature vector. Based on this vector and the IM_{ID} the interval mapping function calculates a biometric hash vector b_{ID} , where interval lengths and offsets provided by IM_{ID} are used to map each of the k statistical features to a corresponding hash value. The biometric hash vector can be used either for cryptographic applications (e.g. key generation) or for biometric verification. In the latter case, the biometric hash vector b_{ID} generated from the currently presented authentication sample D_{ID} is compared against the reference hash vector $b_{ref ID}$ of the claimed identity ID , which in this case needs to be stored as additional information during the enrollment process. The classification can then be performed for example by some distance measurement and comparison to a given threshold T . On the other hand, for verification based on cryptographic hashes (e.g. message authentication codes, MAC) the reference hash and the hash generated for the currently presented data have to be identical, if and only if the hashes generated based on identical data.

In this paper we study the performance of the Biometric Hash algorithm with regard to both, verification mode and hash generation mode, based on different setups, i.e. four different semantics and pair wise multi-semantic fusion.

3 NEW PERFORMANCE MEASURES FOR BIOMETRIC HASHING

Based on the biometric data obtained, a hash generation method aims to generate identical hashes from data of the same person and/or different hashes from data of different users, respectively. In order to provide a measure for the degree of the reproducibility and/or false generation of such hashes, we suggest the Hamming Distance ([10]) as already shown in [9] and [7]. In context of the comparison of two biometric hashes b and b' , the Hamming Distance measure determines the number of positions, where the two hashes are different and returns a value between 0 and the number of elements. In equation (1), b_i and b'_i are the corresponding elements of vectors b and b' at index i . The component-wise comparison of b_i and b'_i yields 0 , if the two elements are equal and 1 otherwise. Then the Hamming Distance between the hashes b and b' is the sum of the results of all single comparisons.

$$hd(b, b') = \sum_{i=0}^{k-1} dist(b_i, b'_i) \quad \text{with } dist(b_i, b'_i) = \begin{cases} 0, & \text{if } b_i = b'_i \\ 1, & \text{else} \end{cases} \quad (1)$$

Derived from the properties of cryptographic hashes, error rates to estimate the performance of biometric hash algorithms should be considered in the reproduction and the collision in addition to *FRR*, *FAR* and *EER*.

In our Hamming Distance based histogram analysis, we compare all generated biometric hashes of each person to each other hash of the same person to calculate the *reproducibility rate* (*RR*). Therefore, a Hamming Distance *hd* of 0 is logged as a match, while any *hd* > 0 is logged as a non-match. Then, the Reproducibility Rate is the quotient of the number of matches by the number of comparisons. The *collision rate* (*CR*) is determined by the comparison of each single person's biometric hashes with the hashes of all other users. For the *CR*, a Hamming Distance of 0 is logged as a collision and all distances higher than 0 are logged as non-collision. The *CR* is calculated by the division of the number of collisions by the number of comparisons. In the ideal case, each comparison between hashes of the same person and semantic should be result in *hd*=0, while the comparison between hashes of any two different persons should yield *hd*>0. In order to refer to reproducibility requirement, the point of interest in the histogram is a Hamming Distance value of 0. This means for *RR*, only the identical reproductions of hashes of the corresponding person are considered, while for the *CR* only identical generations of hashes of non identical persons are examined. However, for the optimization process of a biometric algorithm, the entire Hamming Distance based *HD* distribution should be taken in consideration.

In order to have an indicator of the trade-off relation between *RR* and *CR*, an additional measure is introduced here: the *collision reproducibility ratio* (*CRR*). It is the result of the division of *CR* by *RR*. Since one aim of biometric hashing is to reproduce hashes of each person with a high degree, while hashes of different persons should be different, the *CRR* should be very small.

4 Multi-semantic hash fusion approach

In this section we present a new biometric fusion strategy based on the pair wise combination of the biometric hash vectors of two semantic classes. In the context of biometric handwriting, semantics are alternative written contents in addition to the signature. Semantics can be based on the additional factors of individuality, creativity and/or secret knowledge, e.g. by using pass phrases, numbers or sketches. In [9], Vielhauer shows that the usage of such alternative contents may lead to similar results as the usage of the signature in context of online handwriting based authentication performance. Based on the number of biometric components involved in the fusion process, Ross et al. differentiate in [11] between the following five scenarios for automatic biometric fusion: multi-sensor, multi-algorithmic, multi-instance, multi-sample and multi-modal systems. Since the fusion proposed in this paper is executed on the feature extraction level in the hash domain based on different semantics, it is called multi-semantic hash fusion. It can be assigned to the multi-instance stage of the scheme suggested in [11].

The first step is the data acquisition of two semantics, which form the input for the second step, the hash generation. In this process step, the statistical feature vector is calculated from raw data of each semantic. Then, biometric hash vectors are derived from the semantics' statistical values, as described in the previous section. Although, the tolerance factor TF used for hash generation is identical for both semantics ($TF=3$), it is also feasible to tune the TF separately in dependency on corresponding semantic to optimize the fusion result. The global tolerance vector TV is determined on globally based on disjoint user sets of the corresponding semantics. Thus, for both, statistical feature vectors and biometric hash vectors, the dimensionality is k . The fusion of the two hashes is the last process step, which is carried out as concatenation of both hashes and leads to a hash vector's dimensionality of $2*k$.

5 EVALUATION

This section firstly describes the test data used in our evaluation. Following, our methodologies are presented, which are used to determine the results of biometric handwriting verification as well as biometric hash generation. Finally, the results for both, verification and hashing are presented and discussed.

5.1 Evaluation database

The entire test set is based on 84 users, each of them having donated 10 handwriting samples for four different semantics (total of 3,440 samples). In our test setup, we use four semantics: The *PIN* is given as a sequence of the five digits '77993'. Using this semantic, the individual style of writing plays a more important role than the content, since all test subject write the same numbers. The semantic *Place* represents the individual answer to the question "Where are you from?", written by each test person. This answer includes individual knowledge in a certain degree which, however, is not absolutely secret. We use the semantic *Pseudonym* as anonymous substitution of the individual signature, due to the fact that most of the test subjects refrained from donating their original signature due to privacy concerns. The *Pseudonym* is a name freely chosen by the writer, which had be trained several times before the acquisition. The freely chosen *Symbol* holds individual creative characteristics and additionally provides a knowledge based component in form of the sketched object.

In order to determine a global tolerance vector TV as hash generation parameter and to carry out the biometric error rate analysis and the Hamming Distance histogram analysis, a training set (hereafter set T) of 15 users and an evaluation set (hereafter set E) of 69 users are extracted from the entire set of 86 persons. Both sets are entirely disjoint with respect to the subjects and structured as follows: From the 10 handwriting samples $D=D_1, \dots, D_{10}$ of each person and each semantic, the first 5 samples D_1, \dots, D_5 are taken to create 5 sets, using a leave-one-out strategy. This means a combination of 5 choose 4, i.e. 5 different sets are created, containing 4 handwriting samples each. Each of the 5 sets is used to create a user dependent interval matrix (IM_{ID}) and consequently, we yield reference data $R_i=(ID, IM_{i,ID})$ with $i=1, \dots, 5$. Based on these interval matrices and the remaining samples D_6, \dots, D_{10} , 5

biometric hashes are created for each user of set T and set E respectively. The determination of the tolerance vector TV is conducted globally, based on all users of set T , whereas the biometric error rate analysis and a Hamming Distance based histogram analysis are carried out on disjoint set E .

5.2 Evaluation methodology

In this paper, we use the equal error rate (EER) to show the verification performance of the reference algorithm in comparison to the reproducibility performance of biometric hashes based on dynamic handwriting. For the latter evaluation, we analyze the Biometric Hash algorithm (see section 2.2) by using the new measurements Reproducibility Rate (RR), Collision Rate (CR) and Collision Reproducibility Ratio (CRR) to compare the reference and current hashes as described in section 3.

Note that for the evaluation of the multi-semantic fusion, we assume that there is no temporal dependence between semantic 1 and semantic 2 (i.e. EER , RR , CR or CRR of fusion of semantic 1 and semantic 2 is equal to EER , RR , CR or CRR of fusion of semantic 2 and semantic 1). Thus, the outcome of the fusion is symmetric with respect to the sequence semantics taken into account, and results to the triangular layout of Table 1 and Table 2.

In our previous work, we optimized the tolerance factors TF for verification as well as for hash generation in a certain degree. We observed, that for verification the best integer TF is 1, while for hash generation $TF=3$ was relatively good. Thus, we use in this initial study these both values for the corresponding evaluations. The hash generation for both applications is also based on a global TV determined on a disjoint set of users per semantic. However, it is also possible to use alternative parameterizations for TF and TV to optimize both, verification and hash generation performance.

5.3 Results

This subsection describes the results of the verification and the hash reproducibility. The corresponding tests are carried out on the single semantics as well as on their pair wise fusion. In tables 1 and 2 the best single results are printed in bold, while the best fusion results for EER , RR , CR and CRR are marked with a gray background.

Biometric error rate analysis

Table 1 shows the results of the biometric error rate analysis. While the second column (*single*) presents the $EERs$ of the individual semantics, the last three columns are showing the pair wise fusion results. The fusion is carried out on the matching score level and is based on a simple mean rule. This strategy weights the scores of the two fusion components involved with the same value (0.5) and summates the results to a final fused score. For the verification, the best single-modal result with respect to the EER is determined for the *Symbol* with $EER=3.199\%$. The worst EER of 4.969% is based on semantic *Pseudonym*. Another observation from Table 1 is that all pair wise fusion combinations improve the results determined by the corresponding

semantics. Here the lowest *EER* of 1.143% is calculated based on the combination of *Place* and *Symbol*.

Table 1. Equal error rates in % per semantic class and their pair wise fusion ($TF=1$)

Semantic	single EER	Multi-semantic fusion		
		Symbol EER	Pseudonym EER	Place EER
PIN	4.763	1.719	2.249	1.982
Place	3.541	1.143	1.632	
Pseudonym	4.969	1.382		
Symbol	3.199	-		

Hamming Distance based histogram analysis

The results of the Hamming Distance based histogram analysis for single semantics as well as for their pair wise fusion are presented in Table 2. In the rows of Table 1 labeled with *RR* the reproducibility rate of genuine hashes by the corresponding genuine users is shown in dependency of the semantic class. The rows labeled with *CR* are showing the collision rate, while the *CRR* rows present the collision reproducibility ratio.

Table 2. Reproducibility and collision rate in % and collision reproducibility ratio for single semantics and pair wise semantic hash fusion ($TF=3$)

Semantic 1	Measurement	single results	Semantic 2		
			Symbol	Pseudonym	Place
PIN	RR	76.580	60.000	55.304	55.536
	CR	5.818	0.346	0.685	1.207
	CRR	0.076	0.006	0.012	0.217
Place	RR	72.116	57.217	52.696	
	CR	5.115	0.319	0.484	
	CRR	0.070	0.006	0.009	
Pseudonym	RR	70.551	56.290		
	CR	4.923	0.223		
	CRR	0.070	0.004		
Symbol	RR	77.101			
	CR	2.392	-		
	CRR	0.031			

As shown in the third column of Table 2, the best reproducibility rate of genuine hashes is calculated for *Symbol* with a *RR* of 77.101%. A similar result is calculated based on the *PIN* with *RR*=76.580%. However, since *PIN* is the given sequence of the digits '77993' written by all persons, the collision rate (*CR*=5.818%) is the highest. Thus, also the collision reproducibility ratio for *PIN* (*CRR*=0.076) is higher than the *CRR*s for the other semantics. From the point of view to choose the semantic having the best ratio between *RR* and *CR*, the semantic *Symbol* should be taken in consideration (*CRR*=0.031).

Since the multi-semantic hash fusion is carried out as simple concatenation (see section 4) of two hashes based on different semantics, the reproducibility of the new fused hash depends only on the individual reproducibility of the two hashes involved. Based on this fact, it is obvious that the *RR* of the fused hashes cannot be higher than the worst individual reproducibility rate of the two hashes used for the fusion. Table 2

shows also the results of the pair wise multi-semantic hash fusion. The intersections of rows and columns of the different semantics are showing the corresponding fusion results for reproducibility rate (RR), collision rate (CR) and collision reproducibility ratio (CRR). As assumed, a general observation is, that the fusion results for the reproducibility rate are worse than the results obtained based on the single semantics (see second column of Table 2). For example, the best fusion result is based on the concatenation of the hashes for *PIN* and *Symbol* where the RR is equal to 60%, while the single results amount 76.58% for *PIN* and 77.101% for *Symbol*, respectively. This corresponds to a relative degradation of approx. 22% in comparison to the best single result determined for the *Symbol*. On the other hand, the collision rates are significantly lower than those of the single semantics involved. Here the relative decline lies between 77% and 90%. The best CR of 0.223% was determined for the fusion of semantics *Pseudonym* and *Symbol*, while the corresponding RR amounts 56.29%. The greatest improvement of the fusion we see in the decrease of the CRR . In case of the best fused RR of 60% the CRR is reduced to one fifth (0.006) of the CRR of the best single result calculated for symbol (0.031). Thus, the fusion may provide the opportunity to reach a higher RR at an acceptable CR .

The results of biometric error rate as well as Hamming Distance based histogram analysis show that there is a dependency between EER and/or RR and CR , and the written content. Based on these results it can be stated, that the choice of a semantic depends on the requirements of the verification and/or hashing application. It can be decided on best equal error rate performance or on best reproducibility, best collision resistance as well as on the best ratio between them.

6 CONCLUSIONS

In this paper, we suggest the analysis of the biometric hash reproducibility and collision rates based on the Hamming Distance, in addition to the typical verification error rates. The reproducibility rate (RR) shows, how is the performance of a hash generation algorithm with respect to generate stable has values for the same persons and the same written content. The collision rate (CR) is a measure for the probability of generation of biometric hashes by non-authentic users. Further, the collision reproducibility ratio (CRR), as third introduced measure, indicates the tradeoff relation between CR and RR . In order to find a suitable working point for a biometric hash generation algorithm for practical applications, one solution can be to minimize the CRR . Further, we have suggested a novel concept in the domain of multi-biometrics: Multi-semantic fusion of biometric hashes generated using different writing contents.

In the experimental evaluation, we have practically shown the feasibility of the new measurements based on online handwriting biometrics. On one side, the evaluation of the multi-semantic hash fusion has shown that the concatenation of two hashes using different semantics leads to a significantly worse reproducibility rate than the individual semantics. Here the best fusion result is calculated for the combination of *PIN* and *Symbol* ($RR=60\%$), while the individual RR s for *PIN* and *Symbol* amount 76.580% and 77.101%, respectively. On the other side, a significant

improvement of the collision rate can be observed. The best CR of 0.223% is determined based on the semantics Pseudonym and *Symbol*. This leads to the best collision reproducibility ratio of the entire evaluation ($CRR=0.004$) and this significantly improved trade-off between RR and CR provides potential for optimized parameterization towards better RR at acceptable CR level.

To do so, the parameterization can be adjusted to any user registered in the database by optimizing user specific tolerance vectors, which are used to calculate the mapping interval of the Biometric Hash algorithm. In order to improve the RR even more, other methods have to be studied, e.g. alternative mapping functions or error correction mechanisms. In this case, one has also to keep track of the expansion of CR as counterpart of RR . Finally, although in this paper we have focused on biometric hashes for handwriting, it appears quite possible to apply the methodology to hashes generated based on other biometric modalities in the future.

Acknowledgements. The work on biometric hashes with regard to verification and reproducibility is partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project WritingPrint).

References

- [1] Dittmann, J., Vielhauer, C.: Encryption Related Issues. Joint COST 2101 and BioSecure Industrial and End-user Committee Workshop on “Smart Cards and Biometrics”, Lausanne (2007)
- [2] Monrose, F., Reiter, M. K., Li, Q., Wetzel, S.: Using Voice to Generate Cryptographic Keys. A Speaker Odyssey, The Speech Recognition Workshop, Crete (2001)
- [3] Vielhauer, C., Steinmetz, R., Mayerhöfer, A.: Biometric Hash based on Statistical Features of Online Signature. Proc. of the Intern. Conf. on Pattern Recognition, Quebec City (2002)
- [4] Sutcu, Y., Sencar, H. T., Memon, N.: A Secure Biometric Authentication Scheme Based on Robust Hashing. Proc. of the 7th workshop on Multimedia and security, ACM Press, New York (2005)
- [5] Tulyakov, S., Farooq, F., Govindaraju, V.: Symmetric Hash Functions for Fingerprint Minutiae. International Workshop on Pattern Recognition for Crime Prevention, Security and Surveillance, Bath (2005)
- [6] Korte, U. et al.: A cryptographic biometric authentication system based on genetic fingerprints, In: Alkassar, Siekmann (Eds.): Sicherheit 2008 - Sicherheit, Schutz und Zuverlässigkeit; Beiträge der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V., Saarbrücken (2008)
- [7] Scheidat, T., Vielhauer, C.: Biometric Hashing for Handwriting: Entropy based feature selection and semantic fusion. In: Proc. of SPIE - Security, Steganography and Watermarking of Multimedia Contents X, IS&T/SPIE Symposium on Electronic Imaging, San Jose, California (2008)
- [8] Bishop, M.: Computer Security: Art and Science. Addison-Wesley, Reading, Mass (2003)
- [9] Vielhauer, C.: Biometric User Authentication for IT Security: From Fundamentals to Handwriting. Springer, New York (2006)
- [10] Hamming, R.W.: Error-detecting and error-correcting codes. Bell System Technical Journal XXVI (2) (1950)
- [11] Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, New York (2006)

Additive Block Coding Schemes for Biometric Authentication with the DNA Data

Vladimir B. Balakirsky, Anahit R. Ghazaryan, A. J. Han Vinck

Institute for Experimental Mathematics, Ellernstr. 29, 45326 Essen, Germany
v_b_balakirsky@rambler.ru, a_ghazaryan@rambler.ru, vinck@iem.uni-due.de

Abstract. To implement a biometric authentication scheme, the templates of a group of people are stored in the database (DB) under the names of these people. Some person presents a name, and the scheme compares the template of this person and the template associated with the claimed person to accept or reject their identity [1]. The templates of people stored in the DB should be protected against attacks for discovery the biometrics and attacks for successful passing through the verification test. The authentication algorithm developed by Juels and Wattenberg [2] is a possible solution to the problem. However, implementations of this algorithm for practical data require generalized versions of the algorithm and their analysis. We introduce a mathematical model for DNA measurements and present such a generalization. Some numerical results illustrate the correction of errors for the DNA measurements of a legitimate user and protection of templates against attacks for successful passing the verification stage by an attacker.

1 An additive block coding scheme

An additive block coding scheme proposed in [2] can be presented as follows (see Figure 1). Let \mathcal{C} be a set consisting of M different binary vectors of length n (a binary code of length n for M messages). The entries of the set \mathcal{C} are called key codewords. One of the key codewords $\mathbf{x} \in \mathcal{C}$ is chosen at random with probability $1/M$. This codeword is added modulo 2 to the binary vector \mathbf{b} generated by a biometrical source, and the vector $\mathbf{y} = \mathbf{x} \oplus \mathbf{b}$ is stored in the DB under the name of the person whose biometrics is expressed by the vector \mathbf{b} . Furthermore, the value of a one-way hash function Hash at the vector \mathbf{x} (a one-to-one function whose value can be easily computed, while the inversion is a difficult problem) is also stored in the DB. Having received another binary vector \mathbf{b}' and the claimed name, the verifier finds the key codeword $\hat{\mathbf{x}} \in \mathcal{C}$ located at the minimum Hamming distance from the vector $\mathbf{z} = \mathbf{y} \oplus \mathbf{b}'$. The basis for the algorithm is the observation

$$\left. \begin{array}{l} \mathbf{y} = \mathbf{x} \oplus \mathbf{b} \\ \mathbf{b}' = \mathbf{b} \oplus \mathbf{e} \end{array} \right\} \Rightarrow \mathbf{x} \oplus \mathbf{e} = \mathbf{z}.$$

In particular, if the number of positions where the vectors \mathbf{b} and \mathbf{b}' differ does not exceed $\lfloor (d_C - 1)/2 \rfloor$, where d_C is the minimum distance of the code \mathcal{C} , then the key codeword used at the enrollment stage will be found. Then $\text{Hash}(\hat{\mathbf{x}})$ is equal to $\text{Hash}(\mathbf{x})$ and the identity claim is accepted. Otherwise, the claim is rejected.

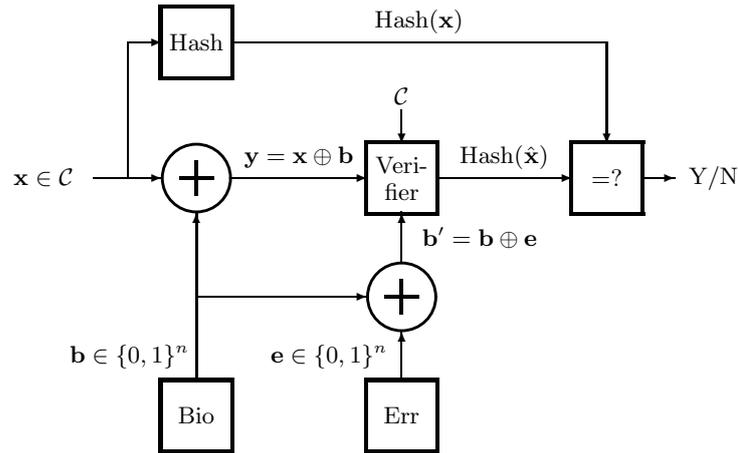


Fig. 1. Verification of a person using an additive block coding scheme with a binary code.

Notice that the verification scheme in Figure 1 can be represented as transmission of the key codeword \mathbf{x} over two parallel channels, because

$$\left. \begin{matrix} \mathbf{y} = \mathbf{x} \oplus \mathbf{b} \\ \mathbf{b}' = \mathbf{b} \oplus \mathbf{e} \end{matrix} \right\} \Rightarrow \begin{cases} \mathbf{x} \oplus \mathbf{b} = \mathbf{y} \\ \mathbf{x} \oplus \mathbf{e} = \mathbf{z}. \end{cases}$$

Thus, we say that the verifier receives a pair of vectors $(\mathbf{x} \oplus \mathbf{b}, \mathbf{x} \oplus \mathbf{e})$ (see Figure 2), while the attacker receives only the first component and the JW decoder analyzes only the second component of that pair. The transformations $\mathbf{x} \rightarrow \mathbf{y}$ and $\mathbf{x} \rightarrow \mathbf{z}$ can be interpreted as transmissions of the key codeword over the biometric and the observation channels, respectively. We will assume that particular binary vectors \mathbf{b} and \mathbf{e} are chosen as the biometric and the observation noise vectors according to the probability distributions (PDs)

$$\left(\Pr_{\text{bio}}\{B = \mathbf{b}\}, \mathbf{b} \in \{0, 1\}^n \right), \left(\Pr_{\text{err}}\{E = \mathbf{e}\}, \mathbf{e} \in \{0, 1\}^n \right).$$

Let \mathbf{x}_{bio} , \mathbf{x}_{err} , and $\mathbf{x}_{\text{bio, err}}$ denote results of the decoding when the vectors \mathbf{y} , \mathbf{z} , and the pair of vectors (\mathbf{y}, \mathbf{z}) are available. One can easily

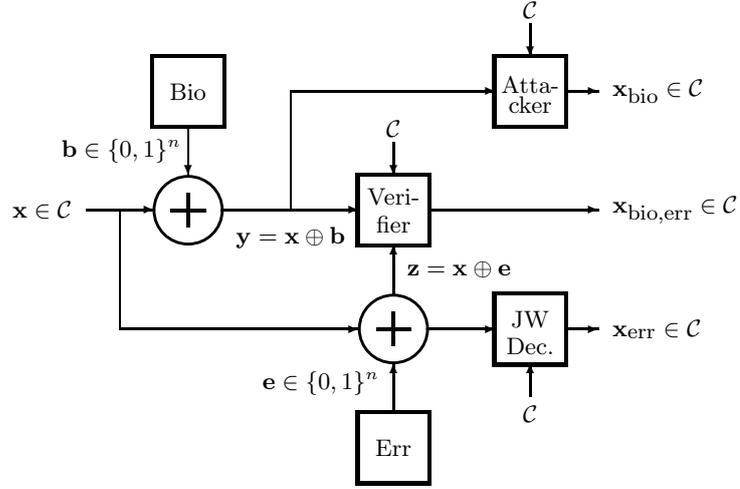


Fig. 2. Representation of the additive block coding as a scheme where a key codeword \mathbf{x} is received under the biometric noise \mathbf{b} and the observation noise \mathbf{e} .

check that the maximum probabilities of correct decoding are attained by the maximum *a posteriori* probability decoding rules, i.e., the optimum estimates of the key codeword satisfy the equalities

$$\Pr_{\text{bio}}\{B = \mathbf{x}_{\text{bio}} \oplus \mathbf{y}\} = \max_{\mathbf{x} \in \mathcal{C}} \Pr_{\text{bio}}\{B = \mathbf{x} \oplus \mathbf{y}\},$$

$$\Pr_{\text{err}}\{E = \mathbf{x}_{\text{err}} \oplus \mathbf{z}\} = \max_{\mathbf{x} \in \mathcal{C}} \Pr_{\text{err}}\{E = \mathbf{x} \oplus \mathbf{z}\},$$

and

$$\Pr_{\text{bio}}\{B = \mathbf{x}_{\text{bio, err}} \oplus \mathbf{y}\} \Pr_{\text{err}}\{E = \mathbf{x}_{\text{bio, err}} \oplus \mathbf{z}\} = \max_{\mathbf{x} \in \mathcal{C}} \left[\Pr_{\text{bio}}\{B = \mathbf{x} \oplus \mathbf{y}\} \Pr_{\text{err}}\{E = \mathbf{x} \oplus \mathbf{z}\} \right].$$

Then the probabilities that the decoded codewords coincide with the transmitted key codewords can be expressed as

$$A_{\text{bio}} = \frac{1}{M} \sum_{\mathbf{y}} \max_{\mathbf{x} \in \mathcal{C}} \Pr_{\text{bio}}\{B = \mathbf{x} \oplus \mathbf{y}\},$$

$$A_{\text{err}} = \frac{1}{M} \sum_{\mathbf{z}} \max_{\mathbf{x} \in \mathcal{C}} \Pr_{\text{err}}\{E = \mathbf{x} \oplus \mathbf{z}\},$$

$$A_{\text{bio, err}} = \frac{1}{M} \sum_{\mathbf{y}, \mathbf{z}} \max_{\mathbf{x} \in \mathcal{C}} \left[\Pr_{\text{bio}}\{B = \mathbf{x} \oplus \mathbf{y}\} \Pr_{\text{err}}\{E = \mathbf{x} \oplus \mathbf{z}\} \right].$$

2 Structure of the DNA data and mathematical model

The most common DNA variations are Short Tandem Repeats (STR): arrays of 5 to 50 copies (repeats) of the same pattern (the motif) of 2 to 6 pairs. As the number of repeats of the motif highly varies among individuals, it can be effectively used for identification of individuals. The human genome contains several 100,000 STR loci, i.e., physical positions in the DNA sequence where an STR is present. An individual variant of an STR is called allele. Alleles are denoted by the number of repeats of the motif. The genotype of a locus comprises both the maternal and the paternal allele. However, without additional information, one cannot determine which allele resides on the paternal or the maternal chromosome. If the measured numbers are equal to each other, then the genotype is called homozygous. Otherwise, it is called heterozygous. The STR measurement errors are usually classified into three groups: (1) *allelic drop-out*, when an allele of a heterozygous genotype is missing, e.g. genotype (7,9) is measured as (7,7); (2) *allelic drop-in*, when in a homozygous genotype, an additional allele is erroneously included, e.g. genotype (10,10) is measured as (10,12); (3) *allelic shift*, when an allele is measured with a wrong repeat number, e.g. genotype (10,12) is measured as (10,13).

The points above can be formalized as follows. Suppose that there are n sources. Let the t -th source generate a pair of integers according to the PD

$$\Pr_{\text{DNA}} \left\{ (A_{t,1}, A_{t,2}) = (a_{t,1}, a_{t,2}) \right\} = \pi_t(a_{t,1})\pi_t(a_{t,2}),$$

where $a_{t,1}, a_{t,2} \in \{c_t, \dots, c_t + k_t - 1\}$ and c_t, k_t are given positive integers. Thus, we assume that $A_{t,1}$ and $A_{t,2}$ are independent random variables that contain information about the number of repeats of the t -th motif in the maternal and the paternal allele. We also assume that $(A_{t,1}, A_{t,2}), t = 1, \dots, n$, are mutually independent pairs of random variables, i.e.,

$$\Pr_{\text{DNA}} \left\{ (A_1, A_2) = (\mathbf{a}_1, \mathbf{a}_2) \right\} = \prod_{t=1}^n \Pr_{\text{DNA}} \left\{ (A_{t,1}, A_{t,2}) = (a_{t,1}, a_{t,2}) \right\},$$

where $A_\ell = (A_{1,\ell}, \dots, A_{n,\ell})$ and $\mathbf{a}_\ell = (a_{1,\ell}, \dots, a_{n,\ell}), \ell = 1, 2$.

Let us fix a $t \in \{1, \dots, n\}$ and denote

$$\mathcal{P}_t \triangleq \left\{ s = (i, j) : i, j \in \{c_t, \dots, c_t + k_t - 1\}, j \geq i \right\}.$$

Then the PD of a pair of random variables

$$S_t \triangleq \left(\min\{A_{t,1}, A_{t,2}\}, \max\{A_{t,1}, A_{t,2}\} \right),$$

which represents the outcome of the t -th measurement, can expressed as

$$\Pr_{\text{DNA}} \left\{ S_t = (i, j) \right\} = \omega_t(i, j),$$

where $\omega_t(i, j) \triangleq \pi_t^2(i)$, if $j = i$, and $\omega_t(i, j) \triangleq 2\pi_t(i)\pi_t(j)$, if $j \neq i$. Denote $\omega_t \triangleq (\omega_t(i, j), (i, j) \in \mathcal{P}_t)$ and

$$\begin{aligned} G(\omega_t) &\triangleq -\log \max_{(i,j) \in \mathcal{P}_t} \omega_t(i, j), \\ H(\omega_t) &\triangleq -\sum_{(i,j) \in \mathcal{P}_t} \omega_t(i, j) \log \omega_t(i, j), \\ p(\omega_t) &\triangleq \sum_{i=c_t}^{c_t+k_t-1} \omega_t(i, i), \\ h(\omega_t) &\triangleq -(1 - p(\omega_t)) \log(1 - p(\omega_t)) - p(\omega_t) \log p(\omega_t). \end{aligned}$$

One can easily see that the best guess of the output of the t -th source is a pair (i_t^*, j_t^*) such that $\omega_t(i_t^*, j_t^*) \geq \omega_t(i, j)$ for all $(i, j) \in \mathcal{P}_t$. Therefore, $2^{-G(\omega_t)}$ is the probability that the guess is correct. The value of $p(\omega_t)$ is the probability that the genotype is homozygous, $H(\omega_t)$ is the entropy of the PD ω_t , and $h(\omega_t)$ is the entropy of the PD $(1 - p(\omega_t), p(\omega_t))$.

Let us assume that $q_t \triangleq |\mathcal{P}_t| = k_t(k_t + 1)/2$ values $\omega_t(i, j), (i, j) \in \mathcal{P}_t$, are different and introduce two transformations of a pair of measurements $(i, j) \in \mathcal{P}_t$. (a) Let $i = j$ imply $\beta(i, j) = 0$ and let $i \neq j$ imply $\beta(i, j) = 1$. (b) Given an integer $q \geq q_t$, let $\beta_q(i, j) = b$ if and only if there are $b - 1$ pairs $(i', j') \in \mathcal{P}_t$ such that $\omega_t(i', j') > \omega_t(i, j)$. In particular, $\beta_q(i_t^*, j_t^*) = 0$.

We will denote the vector of measurements available to the scheme at the enrollment stage by $\mathbf{s} = ((i_1, j_1), \dots, (i_n, j_n))$. The transformations of this vector will be denoted by $\beta(\mathbf{s}) = (\beta(i_1, j_1), \dots, \beta(i_n, j_n))$ and $\beta_q(\mathbf{s}) = (\beta_q(i_1, j_1), \dots, \beta_q(i_n, j_n))$. Similar notations will be used for the vector $\mathbf{s}' = ((i'_1, j'_1), \dots, (i'_n, j'_n))$ available to the scheme at the verification stage.

Example (the quantities below describe the **TH01** allele in Table 1). Let $c_t = 6, k_t = 4$, and $(\pi(6), \dots, \pi(9)) = (0.23, 0.19, 0.09, 0.49)$. Then

$$\left[\pi_t(i)\pi_t(j) \right]_{i,j=6,\dots,9} = \begin{array}{c|c|c|c} & j=6 & j=7 & j=8 & j=9 \\ \hline i=6 & .0529 & .0437 & .0207 & .1127 \\ i=7 & .0437 & .0361 & .0171 & .0931 \\ i=8 & .0207 & .0171 & .0081 & .0441 \\ i=9 & .1127 & .0931 & .0441 & .2401 \end{array}$$

To construct the PD ω_t , we transform this matrix to the right triangular matrix below. The entries above the diagonal are doubled, and the entries below the diagonal are replaced with the zeroes. The sum of all entries of the i -th row is equal to the probability that $\min\{A_{t,1}, A_{t,2}\} = i$ and the sum of all entries of the j -th column is equal to the probability that $\max\{A_{t,1}, A_{t,2}\} = j$ (these sums are denoted by $\omega_{t,\min}(i)$ and $\omega_{t,\max}(j)$),

$$\left[\omega_t(i, j) \right]_{\substack{i, j=6, \dots, 9 \\ j \geq i}} = \begin{array}{c|cccc|c} & j=6 & j=7 & j=8 & j=9 & \omega_{t,\min}(i) \\ \hline i=6 & .0529 & .0874 & .0414 & .2254 & .4071 \\ i=7 & & .0361 & .0342 & .1862 & .2565 \\ i=8 & & & .0081 & .0882 & .0963 \\ i=9 & & & & .2401 & .2401 \\ \hline \omega_{t,\max}(j) & .0529 & .1235 & .0837 & .7399 & \end{array}$$

Reading the entries of this matrix in the decreasing order of their values brings the following table,

i, j	9, 9	6, 9	7, 9	8, 9	6, 7	6, 6	6, 8	7, 7	7, 8	8, 8
$\beta(i, j)$	1	0	0	0	0	1	0	1	0	1
$\beta_q(i, j)$	0	1	2	3	4	5	6	7	8	9
$\omega_t(i, j)$.2401	.2254	.1862	.0882	.0874	.0529	.0414	.0361	.0342	.0081
$G(\omega_t)$	- log .2401 = 2.07									
$p(\omega_t)$.2401 + .0529 + .0361 + .0081 = .3372									

Some parameters of the PDs that were under considerations in the BioKey-STR project [3] are given in Table 1. We conclude that results of the DNA measurements can be represented by a binary vector of length $\lceil \log(q_1 \dots q_n) \rceil = 129$ bits. However the PD over these vectors is non-uniform and (roughly speaking) only 109 bits carry information about the measurements. If an attacker is supposed to guess this vector, then the best guess is the vector of pairs $\mathbf{s}^* = ((i_1^*, j_1^*), \dots, (i_n^*, j_n^*))$. By the construction of the β_q transformation, $\beta_q(\mathbf{s}^*)$ is the all-zero vector. The probability that the guess is correct is equal to $2^{-76.8}$. If the vector of n pairs of integers is transformed to a binary vector of length n containing ones at positions where the genotype is homozygous, then the expected weight of the vector can be computed as $p(\omega_1) + \dots + p(\omega_n) = 7.01$, because the weight is the sum of n independent binary random variables where the t -th variable takes value 1 with probability $p(\omega_t)$. The difference between the entropies $H(\omega_t) - h(\omega_t)$ characterizes the loss of data for the β transformation of presented measurements.

Table 1. Some characteristics of the PDs $\omega_1, \dots, \omega_n$ that describe the DNA measurements for $n = 28$.

t	Name	$\log q_t$	$H(\omega_t)$	$G(\omega_t)$	$p(\omega_t)$	$h(\omega_t)$
1	D8S1179	4.39	4.08	3.01	0.20	0.73
2	D3S1358	3.91	3.71	2.87	0.22	0.76
3	VWA	4.39	4.13	3.12	0.19	0.71
4	D7S820	4.39	4.07	3.25	0.19	0.71
5	ACTBP2	7.71	7.43	6.13	0.06	0.32
6	D7S820	4.81	4.24	3.31	0.19	0.69
7	FGA	5.49	4.92	3.54	0.15	0.61
8	D21S11	4.81	4.13	3.01	0.20	0.73
9	D18S51	5.78	5.28	4.43	0.13	0.55
10	D19S433	4.39	3.59	2.33	0.26	0.82
11	D13S317	4.81	4.15	2.56	0.22	0.75
12	TH01	3.32	2.85	2.07	0.34	0.92
13	D2S138	6.04	5.60	4.23	0.12	0.52
14	D16S539	4.81	3.78	2.25	0.25	0.81
15	D5S818	3.91	3.11	1.81	0.31	0.89
16	TPOX	3.91	2.91	1.79	0.37	0.95
17	CF1PO	3.91	3.16	2.16	0.28	0.86
18	D8S1179	5.49	4.49	3.15	0.19	0.69
19	VWA-1	4.39	4.13	3.12	0.19	0.71
20	PentaD	5.17	4.32	3.13	0.19	0.70
21	PentaE	6.91	5.87	4.02	0.11	0.51
22	DYS390	4.39	3.24	2.06	0.30	0.88
23	DYS429	3.91	2.97	1.78	0.33	0.91
24	DYS437	2.58	2.26	1.58	0.40	0.97
25	DYS391	3.32	1.90	1.11	0.47	1.00
26	DYS385	5.17	3.61	1.72	0.34	0.93
27	DYS389I	2.58	2.01	1.18	0.50	1.00
28	DYS389II	3.91	3.14	2.04	0.31	0.89
	Σ	128.6	109.1	76.8	7.01	21.5

3 Verification of a person using the DNA measurements

Additive block coding schemes are oriented to the correction of certain types of measurement errors with simultaneous hiding biometric data from an attacker. If only the allelic drop-out/in errors are possible, then correction of errors means the transformation of the binary vector $\beta(\mathbf{s}')$ to the binary vector $\beta(\mathbf{s})$, where \mathbf{s} and \mathbf{s}' are biometric vectors presented to the scheme at the enrollment and the verification stages, respectively. This procedure can be organized using an additive block coding scheme with a binary code of length n . However, the β transformation brings an essential loss of input data, and the verifier cannot make a reliable acceptance decision.

Notice that the β_q transformation is lossless. We propose the use of an additive block coding scheme with a q -ary code \mathcal{C}_q , where q is chosen in such a way that $q_1, \dots, q_n \leq q$. All the vectors in Figures 1, 2 become q -ary vectors, and \oplus has to be understood as the component-wise addition modulo q . To distinguish between these vectors and binary vectors, we attach the index q and introduce the following translation to parallel channels:

$$\left. \begin{array}{l} \mathbf{y}_q = \mathbf{x}_q \oplus \mathbf{b}_q \\ \mathbf{b}'_q = \mathbf{b}_q \oplus \mathbf{e}_q \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \mathbf{x}_q \oplus \mathbf{b}_q = \mathbf{y}_q \\ \mathbf{x}_q \ominus \mathbf{e}_q = \mathbf{z}_q \end{array} \right.$$

where $\mathbf{z}_q = \mathbf{y}_q \ominus \mathbf{b}'_q$ and \ominus denotes the component-wise difference modulo q . Our data processing algorithm is presented below.

Preprocessing. Assign a binary code \mathcal{C} for M messages and a q -ary code \mathcal{C}_q for M_q messages. Both codes have length n .

Enrollment (input data are specified by the vector \mathbf{s}).

- (0) Construct the vectors $\beta(\mathbf{s})$ and $\beta_q(\mathbf{s})$.
- (1) Choose a binary key codeword $\mathbf{x} \in \mathcal{C}$. Store $\text{Hash}(\mathbf{x})$ and $\mathbf{y} = \mathbf{x} \oplus \beta(\mathbf{s})$ in the DB.
- (2) Choose a q -ary key codeword $\mathbf{x}_q \in \mathcal{C}_q$. Store $\text{Hash}(\mathbf{x}_q)$ and $\mathbf{y}_q = \mathbf{x}_q \oplus \beta_q(\mathbf{s})$ in the DB.

Verification (input data are specified by the vector \mathbf{s}' and content of the DB).

- (0) Construct the vectors $\beta(\mathbf{s}')$ and $\beta_q(\mathbf{s}')$.
- (1) Consider $(\mathbf{y}, \beta(\mathbf{s}') \oplus \mathbf{y})$ as the pair of received words and decode the binary key codeword as $\hat{\mathbf{x}}$. If $\text{Hash}(\hat{\mathbf{x}}) \neq \text{Hash}(\mathbf{x})$, then output “No” and terminate.
- (2) Consider $(\mathbf{y}_q, \beta_q(\mathbf{s}') \oplus \mathbf{y}_q)$ as the pair of received words and decode the q -ary key codeword as $\hat{\mathbf{x}}_q$. If $\text{Hash}(\hat{\mathbf{x}}_q) \neq \text{Hash}(\mathbf{x}_q)$, then output “No”. Otherwise, output “Yes”.

The formal description of biometric sources for the 1-st and the 2-nd steps are as follows: for all $\mathbf{b} \in \{0, 1\}^n$ and $\mathbf{b}_q \in \{0, \dots, q-1\}^n$,

$$\Pr_{\text{bio}}\{B = \mathbf{b}\} = \prod_{t=1}^n \Pr_{\text{DNA}}\{\beta(S_t) = b_t\},$$

$$\Pr_{\text{bio},q}\{B_q = \mathbf{b}_q\} = \prod_{t=1}^n \Pr_{\text{DNA}}\{\beta_q(S_t) = b_{t,q}\}.$$

Suppose that the noise of observations is specified is such a way that, for all $\mathbf{e} \in \{0, 1\}^n$ and $\mathbf{e}_q \in \{0, \dots, q-1\}^n$,

$$\Pr_{\text{err}}\{E = \mathbf{e}\} = \prod_{t=1}^n \begin{cases} 1 - \varepsilon, & \text{if } e_t = 0, \\ \varepsilon, & \text{if } e_t = 1, \end{cases}$$

$$\Pr_{\text{err},q}\{E = \mathbf{e}_q\} = \prod_{t=1}^n \begin{cases} 1 - \varepsilon_q, & \text{if } e_{t,q} = 0, \\ \varepsilon_q/(q-1), & \text{if } e_{t,q} \in \{1, \dots, q-1\}, \end{cases}$$

where ε and ε_q are given.

Let us estimate the decoding error probability at the output of the JW decoders. One can easily see that if the decoder tries to find a key codeword at distance at most $\lfloor (d_C - 1)/2 \rfloor$ from the received vector \mathbf{y} and outputs an error when it is not possible, then the probability of correct decoding is expressed as

$$\hat{\Lambda}_{\text{err}}(\varepsilon) = \sum_{\nu=0}^{\lfloor (d_C-1)/2 \rfloor} \binom{n}{\nu} (1 - \varepsilon)^{n-\nu} \varepsilon^\nu.$$

The decoding at the 2-nd step can be organized as a procedure that depends on the results of the 1-st step. Namely, the decoder can replace symbols of the vector \mathbf{y}_q located at positions where the vector $\hat{\mathbf{e}} = \mathbf{y} \oplus \hat{\mathbf{x}}$ contains 1's with erasures and decode the resulting vector $\hat{\mathbf{y}}_q$. One can easily see that an estimate of the probability of correct decoding can be expressed as

$$\hat{\Lambda}_{\text{err}}^*(\varepsilon, \varepsilon_q) = \sum_{\nu=0}^{\lfloor (d_C-1)/2 \rfloor} \binom{n}{\nu} (1 - \varepsilon)^{n-\nu} \varepsilon^\nu \hat{\Lambda}_{\text{err},q}(\varepsilon_q | \text{wt}(\hat{\mathbf{e}})),$$

where

$$\hat{\Lambda}_{\text{err},q}(\varepsilon_q | \text{wt}(\hat{\mathbf{e}})) \triangleq \sum_{\tau=0}^{\lfloor (d_{C_q} - \text{wt}(\hat{\mathbf{e}}) - 1)/2 \rfloor} \binom{n - \text{wt}(\hat{\mathbf{e}})}{\tau} (1 - \varepsilon_q)^{n - \text{wt}(\hat{\mathbf{e}}) - \tau} \varepsilon_q^\tau$$

Table 2. Estimates of the decoding error probability for $n = 28$ and $d_{C_q} = 5$.

ε	$1 - \hat{A}_{\text{err}}(\varepsilon)$			$1 - \hat{A}_{\text{err}}^*(\varepsilon, \varepsilon_q = .001)$		
	$d_C = 5$	$d_C = 7$	$d_C = 9$	$d_C = 5$	$d_C = 7$	$d_C = 9$
.001	3.2e-06	2.0e-08	9.6e-11	1.6e-05	1.3e-05	1.3e-05
.002	2.5e-05	3.2e-07	3.0e-09	4.7e-05	2.3e-05	2.2e-05
.003	8.4e-05	1.6e-06	2.3e-08	1.1e-04	3.4e-05	3.3e-05
.004	1.9e-04	4.9e-06	9.3e-08	2.3e-04	4.9e-05	4.4e-05
.005	3.7e-04	1.2e-05	2.8e-07	4.2e-04	6.8e-05	5.7e-05

is the estimate of the probability of correct conditional decoding at the 2-nd step. Some numerical results are given in Table 2.

Considerations presented in [4] show that the performance of the verifier, who analyzes transmitted key codeword both under the biometric and the observation noise, corresponds to the performance of the JW decoder for the channel having crossover probability $\varepsilon' = 2\varepsilon/3$, i.e., $\hat{A}_{\text{bio, err}}(\mathbf{p}, \varepsilon) = \hat{A}_{\text{err}}(2\varepsilon/3)$. The value of parameter ε that can be of interest for practical systems is $\varepsilon = 0.005$, and the corresponding values of the decoding error probabilities are given in Table 2 in bold font.

We can also prove the following upper bound on the probability of correct decoding by the attacker,

$$\hat{A}_{\text{bio}}(\mathbf{p}) \leq \frac{2^n}{M} \cdot \frac{q^n}{M_q} \max_{\mathbf{s}} \Pr_{\text{DNA}} \{ S = \mathbf{s} \}.$$

In particular, if \mathcal{C} is the code for $M = 2^{14}$ messages having the minimum distance 8 and \mathcal{C}_8 is the Reed–Solomon code over $GF(2^8)$ for $M_8 = (2^8)^{24}$ messages having the minimum distance 5, then $\hat{A}_{\text{bio}}(\mathbf{p})$ is equal to $2^{-14} 2^{-8(28-24)} 2^{-76.8} = 2^{-30.8}$.

A more detailed discussion of the implementation issues will be presented in another paper.

References

- [1] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, A. W. Senior, *Guide to Biometrics*. NY: Springer, 2004.
- [2] A. Juels, M. Wattenberg, “A fuzzy commitment scheme,” *Proc. ACM Conf. Computer and Communication Security*, 1999.
- [3] U. Korte, M. Krawczak, J. Merkle, R. Plaga, M. Niesing, C. Tiemann, A. J. Han Vinck, U. Martini, “A cryptographic biometric authentication system based on genetic fingerprints”. Presented at Sicherheit–2008 and available at the Web.
- [4] V. B. Balakirsky, A. R. Ghazaryan, and A. J. Han Vinck, “Performance of additive block coding schemes oriented to biometric authentication”. *Proc. 29th Symposium on Information Theory in the Benelux*, Leuven, Belgium, 2008.

Template Protection for On-line Signature-based Recognition Systems

Emanuele Maiorana, Patrizio Campisi, and Alessandro Neri

Dip. Elettronica Applicata
Università degli Studi Roma Tre
via Della Vasca Navale 84, I-00146, Roma, Italy
Tel:+39.06.5517.7064, Fax:+39.06.5517.7026
e-mail: (campisi, maiorana, neri)@uniroma3.it

Abstract. Security and privacy issues are considered as two of the major concerns related to the use of biometric data for authentication purposes. In this paper we propose two different approaches for the protection of on-line signature biometric templates. In the first one, cryptographic techniques are employed to protect signature features, making impossible to derive the original biometrics from the stored templates. In the second one, data hiding techniques are used to design a security scalable authentication system, embedding some dynamic signature features into a static representation of the signature itself. Extensive experimental results are provided to testify the effectiveness of the presented protection methods.

1 Introduction

The most emerging technology for people authentication is biometrics. Being based on strictly personal traits, much more difficult to be forgotten, stolen, or forged than traditional data employed for authentication, like passwords or ID cards, biometric-based recognition systems typically guarantee improved comfort and security for their users. Unfortunately, the use of biometric data involves various risks not affecting other approaches: significant privacy concerns arise since biometrics can be used, in a fraudulent scenario, to treat the user anonymity which must be guaranteed in many real life situations [1]. Moreover, in a scenario where biometrics can be used to grant physical or logical access, security issues regarding the whole biometric system become of paramount importance. Therefore, when designing a biometric-based recognition system, the issues deriving from security and privacy concerns have necessarily to be carefully considered, trying to provide countermeasures to the possible attacks that can be perpetrated at the vulnerable points of the system, detailed in [2].

In this paper we focus on the signature templates security, presenting two different methods for the protection of the considered biometric data. In Section 1.1 some approaches already proposed for the protection of biometrics are discussed. Our methods are presented in Section 2, where a user-adaptive fuzzy commitment scheme is designed with application to on-line signature based authentication, and Section 3, where a different perspective is taken, employing data hiding techniques to design a security scalable authentication system. An extensive discussion on the performances of the proposed systems is given in Section 4, while the conclusions are finally drawn in Section 5.

1.1 Biometric Template security: state of the art

Different solutions have been investigated to secure biometric templates. Among them, data hiding techniques can be implemented to protect or authenticate biometric data, according to two different possible scenarios: one where the information to hide is of primary concern, in which case we speak about *steganography*, and the other where the host data is of primary concern and the mark is used to validate the host data itself, in which case we talk about *watermarking*. The use of data hiding techniques for biometrics protection has already been proposed in [3, 4], among the others. Although cryptography and data hiding can be properly used to generate secure template, the most promising approaches for biometric template protection consist in the implementation of what has been called *cancelable biometrics*. Originally introduced in [2], it can be roughly described as the application of an intentional and repeatable modification to the original biometric template, able to guarantee the properties or renewability and security for the generated templates. A classification of the proposed protection methods have been presented in [5], comprising two macro-categories, referred to as *Biometric Cryptosystem* and *Feature Transformation* approach. Biometric cryptosystems typically employ binary keys in order to secure the biometric templates. This category can be further divided in *key binding* systems, where the helper data are obtained by binding a key with the biometric template [6], and *key generating* systems, where both the helper data and the cryptographic key are directly generated from the biometric template [7, 8]. In a feature transformation approach, a transformation function is applied to the biometric template, and the desired cancelable biometrics are given by the transformed versions of the original data. It is possible to distinguish between *salting* approaches, where the employed transformation functions are invertible [9], and *non-invertible transform* approaches, where a one-way function is applied to the templates [10]. Considering on-line signature protection, the first proposed (key generation) approaches have been in [11] and [12]. In [13] an adaptation of the fuzzy vault [8] is proposed. Also the fuzzy commitment [7] (whose most established implementation is known as Helper Data System [14]) has been employed to provide security to the features extracted from an on-line signature [15]. A comprehensive survey on signature template protection can be found in [16].

2 Signature-based user adaptive fuzzy commitment

In this Section a key binding scheme for the protection of the on-line signature templates protection is presented. Basically, it is based on Juels' proposal of fuzzy commitment using error correcting codes [7]. The proposed approach is twofold, allowing the system both to manage cancelable biometrics and to handle the intra-class variability exhibited by biometric signatures.

2.1 Enrollment stage

The proposed enrollment scheme is presented in Figure 1. During the enrollment phase a number I of signatures are recorded for each subject s . The 95 features detailed in [17] are then extracted from each signature, and collected in the

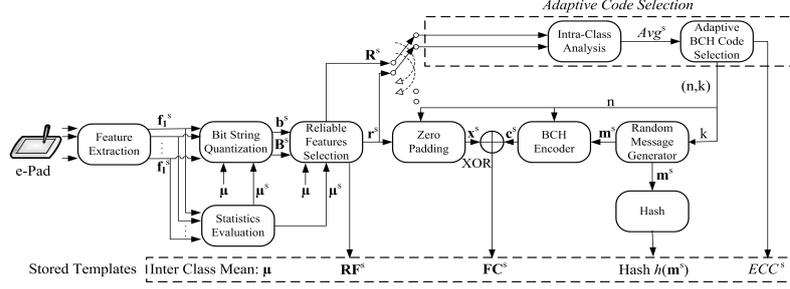


Fig. 1. Signature-based fuzzy commitment: enrollment scheme. The acquired data are analyzed, quantized and summed to error correcting codes.

vectors \mathbf{f}_i^s , $i = 1, \dots, I$. The intra-class $\boldsymbol{\mu}^s$ and the inter-class $\boldsymbol{\mu}$ vector mean are then estimated as $\boldsymbol{\mu}^s = 1/I \sum_{i=1}^I \mathbf{f}_i^s$, where $\boldsymbol{\mu} = 1/S \sum_{s=1}^S \boldsymbol{\mu}^s$, being S the number of enrolled subjects. From the I signatures acquired from the user s , a binary vector \mathbf{b}^s , representative of the considered P features, is then obtained applying, to the intra-class mean vector $\boldsymbol{\mu}^s$, the vector $\boldsymbol{\mu}$ as a threshold. A selection of the relevant features is then performed: only subjects' most reliable features are selected, thus counteracting the potential instability, for the single user, of the feature vector components. In the process of defining a reliable feature selection, for each user s , the enrolled features vectors \mathbf{f}_i^s , with $i = 1, \dots, I$, are binarized by comparisons with the inter-class mean $\boldsymbol{\mu}$ and collected as row vectors in a binary matrix \mathbf{B}^s , with I (signature samples) rows and P (features) columns. Then, the reliability $\mathbf{L}_1^s[p]$ of the p -th feature is defined as follows:

$$\mathbf{L}_1^s[p] = 1 - \frac{\sum_{i=1}^I (\mathbf{B}^s[i, p] \oplus \mathbf{b}^s[p])}{I}, \quad p = 1, \dots, P, \quad (1)$$

where \oplus represents the XOR operation. According to this measure, components with a high reliability possess a high discrimination capability. In order to further discriminate among the available features, we introduce a second level of feature screening, according to the following reliability measure:

$$\mathbf{L}_2^s[p] = \frac{|\boldsymbol{\mu}[p] - \boldsymbol{\mu}^s[p]|}{\boldsymbol{\sigma}^s[p]}, \quad p = 1, \dots, P, \quad (2)$$

with $\boldsymbol{\sigma}^s[p] = \sqrt{\frac{1}{I-1} \sum_{i=1}^I [\mathbf{f}_i^s[p] - \boldsymbol{\mu}^s[p]]^2}$ being the standard deviation of the p -th feature of subject s . A higher discriminating power is thus trusted to features with a larger difference between $\boldsymbol{\mu}^s[p]$ and $\boldsymbol{\mu}[p]$, relative to the standard deviation $\boldsymbol{\sigma}^s[p]$. After the application of the reliability metrics to \mathbf{b}^s , we end up with the binary feature vector \mathbf{r}^s containing the P' most reliable components of \mathbf{b}^s . The indexes of the most reliable feature for the user s are collected in \mathbf{RF}^s .

In order to achieve both template protection and renewability, our scheme uses error correcting codes (BCH codes) [18]. In this paper, we propose an authentication method that provides also adaptability to the user signature variability: this is achieved by choosing the BCH code and its ECC in such a way

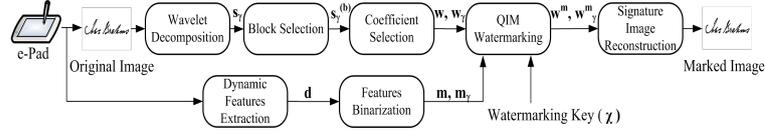


Fig. 2. Security-scalable signature-based authentication system using data hiding. Proposed enrollment scheme.

that, for users characterized by a high intra-class variability, codes with higher error correction capabilities are selected. Therefore, in the enrollment stage, an intra-class analysis is performed as follows: once the P' reliable features are selected, the matrix \mathbf{R}^s , having I rows and P' columns, is obtained from \mathbf{B}^s dropping the non-reliable features. Then, the Hamming distances D_i^s , with $i = 1, \dots, I$, between any rows of \mathbf{R}^s and the representative vector \mathbf{r}^s are evaluated. The average Avg^s of the D_i^s values, $Avg^s = 1/I \sum_{i=1}^I D_i^s$, is then used to characterize the intra-class variability of the user s . Specifically, the BCH code whose ECC is equal to the nearest integer of $[Avg^s + \Delta_{ECC}]$, where Δ_{ECC} is a system parameter common to all the enrolled users, is chosen.

Finally, the binary vector \mathbf{r}^s is zero padded in order to reach the same length n of the selected BCH codewords, resulting in the vector \mathbf{x}^s . The fuzzy commitment \mathbf{FC}^s is then generated using a codeword \mathbf{c}^s obtained from the encoding of a random message \mathbf{m}^s : $\mathbf{FC}^s = FC(\mathbf{x}^s, \mathbf{c}^s) = \mathbf{x}^s \oplus \mathbf{c}^s$. A hashed version $h(\mathbf{m}^s)$ of \mathbf{m}^s , created using the SHA-256 algorithm is eventually stored.

2.2 Authentication stage

The authentication phase follows the same steps as the enrollment stage. When a subject claims his identity, he provides his signature, which is converted in the features vector $\tilde{\mathbf{f}}^s$. Then the quantization is done using the inter-class mean μ , thus obtaining $\tilde{\mathbf{b}}^s$. The reliable features $\tilde{\mathbf{r}}^s$ are selected using $\mathbf{R}\mathbf{F}^s$, and later extended using zero padding, generating $\tilde{\mathbf{x}}^s$. A binary vector $\tilde{\mathbf{c}}^s$, representing a possibly corrupted BCH codeword, results from the XOR operation $\tilde{\mathbf{c}}^s = \tilde{\mathbf{x}}^s \oplus \mathbf{FC}^s$. The BCH decoder is selected depending on the encoder used in enrollment, obtaining $\tilde{\mathbf{m}}^s$ from $\tilde{\mathbf{c}}^s$. Finally, the SHA-256 hashed version $h(\tilde{\mathbf{m}}^s)$ is compared to $h(\mathbf{m}^s)$: if both values are identical the subject is authenticated.

3 Signature Recognition System using data hiding

In this Section we propose a signature-based biometric system, where data hiding is applied to signature images in order to hide and keep secret some dynamic signature features (which can not be derived from a still signature image) in a static representation of the signature itself. The marked images can be used for user authentication, letting their static characteristics being analyzed by automatic algorithms or security attendants. When needed, the hidden dynamic features can be extracted and used to enforce the authentication procedure. Specifically, the fusion of static and dynamic signature features is performed when a high security level is requested.

Index	Description	Index	Description
1	Sample Count	52-54	Height, Width and Aspect Ratio
2-3	X and Y Area	55-57	Minimum, Mean and Maximum X Position
4-15	Mean Pressure 12-segm.	58-60	Minimum, Mean and Maximum Y Position
16-27	Sample Count 12-segm.	61-65	Statistical Moment $M_{1,1}, M_{1,2}, M_{2,1}, M_{0,3}, M_{3,0}$
28-51	X and Y Area 12-segm.	66-68	Minimum, Mean and Maximum Pressure Value

Table 1. Static features extracted from each signature image.

3.1 Enrollment Stage

The enrollment procedure of the proposed security-scalable signature-based authentication system is sketched in Figure 2, and detailed in the following. It is worth pointing out that we use the pressure values of the signature as the host signal where to embed the watermark, thus achieving a higher discriminative capability for the considered signature images, with respect to the simple binary signature images employed by conventional methods.

Both some dynamic features to be embedded in the signature image, and some static features which will be used to perform the first level of user authentication, are extracted during enrollment. For a given user u , the 68 static features detailed in Table 1 are extracted from each of the I acquired signatures.

We consider both global (the first 20) and local features (the last 48), calculated by dividing each signature image, of dimension 720×1440 pixels, in 12 equal-sized rectangular segments [19]. Among the I signatures acquired for the user u , a representative signature is selected to be the host image where to embed the selected user's dynamic features. This is accomplished taking the signature image whose static features of Table 1 are the closest (in an Euclidean sense) to the mean estimated from the I acquired signature.

The chosen pressure image $s[i, j]$ undergoes a two-level wavelet decomposition. The second level subbands, $s_{2LL}[i, j]$, $s_{2HL}[i, j]$, $s_{2LH}[i, j]$, and $s_{2HH}[i, j]$, which represent the approximation and the horizontal, the vertical, and the diagonal detail subbands respectively, are selected for the embedding. Being signature images typically sparse images, the subbands $s_\gamma[i, j]$, with $\gamma \in \Gamma = \{2LL, 2HL, 2LH, 2HH\}$, are then decomposed into blocks of $P \times P$ pixels, in order to identify the proper areas where the watermark has to be embedded: having indicated with $s_\gamma^{(b)}[i, j]$ the generic b -th block extracted from the subband γ , it is selected for watermark embedding if its energy is greater than a fixed threshold T_E , that is, if the block contains a meaningful fragment of the signature. The selected blocks are then projected in the Radon-Discrete Cosine Transform (R-DCT) domain introduced in [4]: this transformation is implemented applying the finite Radon Transform (FRAT) [20] to each considered block, and then performing on each FRAT projection sequence an energy compaction by means of the DCT. Formally, the R-DCT of a selected blocks $s_\gamma^{(b)}[i, j]$ can be written as:

$$c_\gamma^{(b)}[k, q] = \omega[l] \sum_{l=0}^{P-1} r_\gamma^{(b)}[k, l] \cos \left[\frac{\pi(2l+1)q}{2P} \right], \quad r_\gamma^{(b)}[k, l] = \frac{1}{\sqrt{P}} \sum_{(i,j) \in L_{k,l}} s_\gamma^{(b)}[i, j] \quad (3)$$

Index	Description	Assigned Bits
1	Number of the Strokes	5
2	Time Duration	7
3	Pen Up/Pen Down Ratio	8
4-5	Number of X and Y Maximums	6 + 6
6-7	Initial and Final X	10 + 10
8-9	Initial and Final Y	10 + 10
10-11	Mean Instantaneous Velocity and Acceleration Direction	10 + 10

Table 2. Dynamic features extracted from each signature.

where $r_\gamma^{(b)}[k, l]$ represents the FRAT [20], and having indicated with $L_{k,l}$ the set of points that form a line. on Z_P^2 .

Among the $P + 1$ available projections, only the sequences associated to the two most energetic direction k_1 and k_2 of each block are selected to be marked. From them, the matrix $\mathbf{M}_\gamma^{(b)}$ is then built taking the N first components from each sequence (without considering the DC coefficients of the projections):

$$\mathbf{M}_\gamma^{(b)} = \begin{pmatrix} c_\gamma^{(b)}[k_1, 1] & c_\gamma^{(b)}[k_1, 2] & \cdots & c_\gamma^{(b)}[k_1, N] \\ c_\gamma^{(b)}[k_2, 1] & c_\gamma^{(b)}[k_2, 2] & \cdots & c_\gamma^{(b)}[k_2, N] \end{pmatrix}. \quad (4)$$

Iterating this procedure for all the B_γ blocks selected from each subband γ , four host vectors \mathbf{w}_γ , where the mark has to be embedded, can be generated, considering the concatenation of the vectors originated by scanning the matrices $\mathbf{M}_\gamma^{(b)}$ column-wise. The watermarks are generated by extracting from each user's signature the dynamic features detailed in Table 2. The mean dynamic features vector is then binarized using the bit depths given in Table 2. The so obtained binary vector, with length equal to 92 bits, is then BCH coded to provide error resilience. We have chosen to use a (127,92) BCH code, which provides an error correction capability (*ECC*) equal to 5 bits. The coded binary vector \mathbf{m} , consisting of 127 bits, is then decomposed into 3 separate marks \mathbf{m}_{2LL} , \mathbf{m}_{2HL} and \mathbf{m}_{2LH} with dimensions equal to 32 bits, and a fourth mark \mathbf{m}_{2HH} with dimension equal to 31 bits. These marks are separately embedded, by means of QIM [21] watermarking, in the corresponding hosts \mathbf{w}_γ , $\gamma \in \Gamma$.

3.2 Authentication Stage

In the authentication stage the user is asked to provide his signature by means of an electronic pad. His prototype signature with the embedded signature dynamic information can be stored either in a centralized database or in a card. When a low-security level is required the authentication is performed on the base of the selected static features only. Otherwise, when a high-security level is needed, also the the dynamic features embedded in the stored signature are extracted and compared with the acquired ones. A Mahalanobis distance is used to match the extracted features vectors, employing the standard deviations, estimated during enrollment, of both static and dynamic features. Moreover, the best recognition rates, as it will be outlined in Section 4.2, can be obtaining from the fusion of both static and dynamic information. This can be accomplished using score

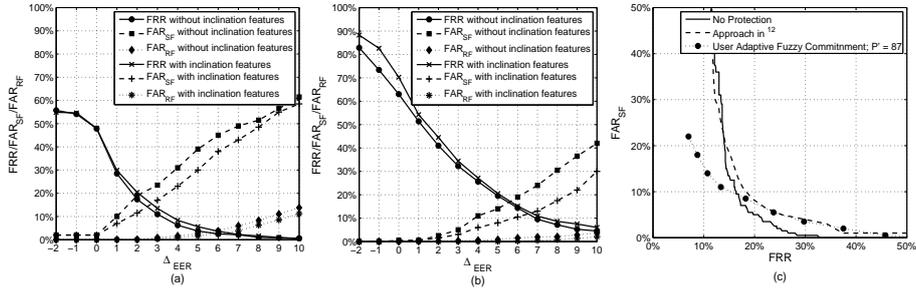


Fig. 3. Proposed fuzzy commitment-based system’s performances. (a): $P' = 50$; (b) $P' = 80$; (c) Comparison between the performances of the adaptive fuzzy commitment, a system without protection, and the one proposed in [12].

fusion techniques [22]. Specifically, we used the *double sigmoid* normalization technique, which is robust to outliers in the score distribution, followed by *sum* fusion technique, thus obtaining a single fused matching score.

4 Experimental Results

In this Section an extensive set of experimental results concerning the performances of the proposed signature-based authentication systems are presented. A database comprising 30 users, from each of which 50 signatures have been acquired during a week time span, has been used to test the effectiveness of the presented approaches. Also a test set of ten skilled forgeries for each subject, created using a training time of ten minutes for each signature whose original was made available to the forger, has been made available.

4.1 Experimental Results: Signature-based fuzzy commitment

In this Section, the recognition performances achievable using the proposed fuzzy commitment-based system for the protection of signature templates are presented. For each user, $I = 10$ signatures have been considered during the enrollment stage. In Figure 3(a) the system performances obtained using the set of features indicated in [17], and considering only the ($P' = 50$) most reliable features for each user, are given. Two different scenarios, one where pen-inclination-dependant features are not considered, and one the whole set with 95 features is considered, are taken into account. In order to show the effectiveness of the proposed feature selection procedure, the system performances achieved when ($P' = 80$) reliable features are also displayed in Figure 3(b). The results are shown with respect to the parameter Δ_{ECC} , used to determine the proper error correction capability for each user. The performances have been assessed in terms of False Rejection Rate (FRR), False Acceptance Rate (FAR) in conditions of skilled forgeries (FAR_{SF}), and FAR in conditions of random forgeries (FAR_{RF}), where the signatures of the users different from the analyzed one are employed as forgeries. The achieved equal error rates (EER)s are approximately 19% (without pen-inclination features) and 16% (with pen-inclination features)

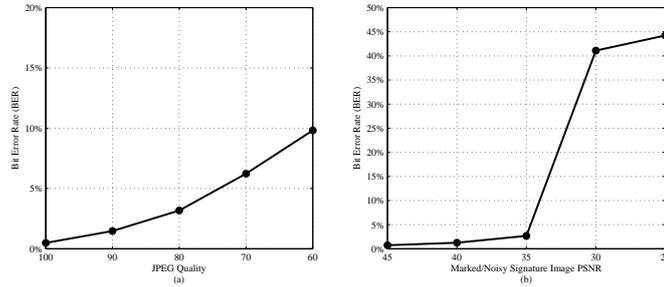


Fig. 4. Mark extraction performances. (a): *BER* vs. JPEG quality level; (b): *BER* vs. marked and noisy image PSNR.

for $P' = 50$, whereas $EER = 16\%$ (without pen-inclination features) and $EER = 12\%$ (with pen-inclination features) for $P' = 80$, considering skilled forgeries.

A performance comparison among the proposed method, the one where no template protection is taken into account, and the one in [12], which also relies on the processing of parametric features extracted from signatures, is also reported in Figure 3(c). Specifically, for the unprotected approach we used a Mahalanobis distance as feature vector matcher. The performances of the method proposed in [12] are very close to those obtainable when no protection is applied. As far as the proposed adaptive scheme is concerned, the obtained ROC curves differentiate with respect to the one obtained when no protection is taken into account: better performances in terms of FRR are obtained (lower value is equal to 7%) making the proposed approach more suitable to forensic applications. Moreover, the best achievable EER is obtained using our adaptive fuzzy commitment approach, and is equal to 12%.

4.2 Experimental Results: Signature-based Authentication System using Data Hiding

The performances regarding this system have been characterized in terms of both the robustness of the employed watermarking method and of the recognition capabilities.

Mark Extraction Performances The performances of the proposed embedding method are evaluated on the basis of the available 1500 genuine signature images. The embedding, detailed in Section 3, is performed using random binary marks of 127 bits which, in our case, represent the BCH encoded dynamic features extracted from the acquired signature. Some attacks, like JPEG compression and additive random Gaussian noise, have been performed on the watermarked signature images for testing the robustness of the proposed embedding methods. The obtained results are displayed in Figure 4, where $P = 10$, $T_E = 5$, and $N = 6$ have been considered as system's parameters. Figure 4(a) shows the obtained bit-error-rate (*BER*) as a function of the JPEG quality of the marked image, while Figure 4(b) shows the *BER* obtained when considering marked images with Gaussian noise added, as a function of the PSNR between the marked

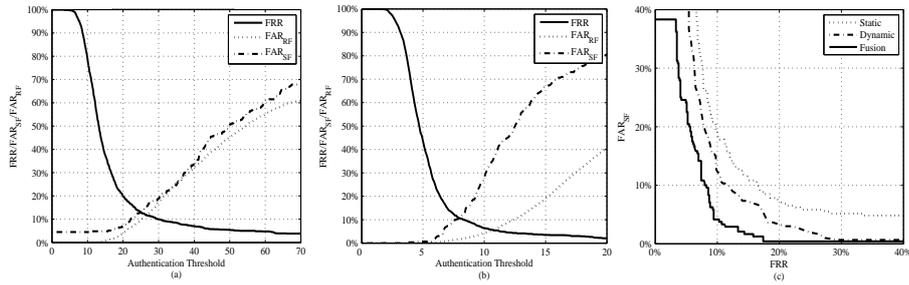


Fig. 5. Proposed data hiding-based recognition system’s performances. (a): static features; (b): dynamic features; (c): individual and combined systems.

and the noisy signature images. As can be seen, the achieved results allow us to properly extract the embedded features (using the error correcting capability of the employed BCH codes) for compression with JPEG quality equal to 80, or PSNR equal to 35.

Recognition Performances In Figure 5 the obtained recognition performances, referred to a case where $I = 10$ signatures have been considered during enrollment, are reported. Figure 5(a) and 5(b) show respectively the performances obtained using only static features, and only dynamic features. In Figure 5(c) the results related to the fusion of static and dynamic features are displayed. All the images we have considered were compressed with a JPEG quality value equal to 90. The embedding is performed using $P = 10$ pixels, $T_E = 5$ and $N = 6$. As it is shown, the equal error rate (EER) achievable using only static features is approximately 15% considering random forgeries, and 17% considering skilled forgeries. The use of dynamic features results in an EER of approximately 10% for random forgeries and 18% for skilled forgeries. Moreover, the performances obtainable from the combined system are better than those of the individual ones, resulting in $EER = 12, 5\%$, considering skilled forgeries.

5 Conclusions

In this paper we present two different approaches to protect a signature biometric template. A user adaptive template protection scheme applied to signature biometrics, which stems from the fuzzy commitment scheme, is proposed. The system is able to provide performances comparable with those achievable by a non-protected system. Moreover, data hiding techniques are also used to design a security scalable authentication system. Specifically, watermarking has been employed to hide some dynamic signature features into a static representation of the signature itself. Experimental results characterizing the system performances in terms of both the achievable authentication capabilities, and the robustness of the implemented watermarking technique, are reported.

References

1. S. Prabhakar, S. Pankanti, A.K. Jain, “Biometric Recognition: Security and Privacy Concerns”, *IEEE Security & Privacy Magazine* 1, pp: 33–42, 2003.

2. N. Ratha, J. H. Connell, R. M. Bolle, "An analysis of minutiae matching strength", *Proc. Int. Conf. Audio and Video-based Biometric Person Authentication*, 2001.
3. A.K. Jain, U. Uludag, "Hiding Biometric Data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 11, pp: 1494–1498, 2003.
4. E. Maiorana, P. Campisi, A. Neri, "Biometric Signature Authentication Using Radon Transform-Based watermarking Techniques", *IEEE Biometric Symposium*, Baltimore, MD, USA, 2007.
5. A.K. Jain, K. Nandakumar, A. Nagar, "Biometric Template Security", *EURASIP Journal on Advances in Sign. Proc., Special Issue on Biometrics*, 2008.
6. Y. Sutcu, Q. Li, N. Memon, "Protecting Biometric Templates with Sketch: Theory and Practice", *IEEE Trans. on Inf. Forensics and Security*, Vol. 2, No. 3, 2007.
7. A. Juels, M. Wattenberg, "A Fuzzy Commitment Scheme", *6th ACM Conf. Computer and Communication Security*, Singapore, pp: 28–36, 1999.
8. A. Juels, M. Sudan, "A Fuzzy Vault Scheme", *Proceedings IEEE on International Symposium on Information Theory*, Lausanne, Switzerland, p. 408, 2002.
9. A.B.J. Teoh, D.C.L. Ngo, A. Goh, "Random Multispace Quantization as an Analytic Mechanism for BioHashing of Biometric and Random Identity Inputs", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 12, 2006.
10. N. Ratha, S. Chikkerur, J. H. Connell, R. M. Bolle, "Generating Cancelable Fingerprint Templates", *IEEE Transactions on PAMI*, Vol. 29, No. 4, 2007.
11. C. Vielhauer, R. Steinmetz, A. Mayerhöfer, "Biometric Hash based on statistical Features of online Signatures", *Int. Conf. on Pattern Recognition*, Vol.1, 2002.
12. H. Feng, C.W. Chan, "Private Key Generation from On-line Handwritten Signatures", *Information Management and Computer Security*, pp: 159–164, 2002.
13. M. Freire-Santos, J. Fierrez-Aguilar, J. Ortega-Garcia, "Cryptographic key generation using handwritten signature", *SPIE, Defense and Security Symposium, Biometric Technologies for Human Identification*, Vol. 6202, pp. 225–231, 2006.
14. M. Van der Veen, T. Kevenaer, G.-J. Schrijen, T.H. Akkermans, F. Zuo, "Face biometrics with renewable templates", in *SPIE Proceedings on Security, Steganography, and Watermarking of Multimedia Contents*, Vol. 6072, 2006.
15. E. Maiorana, P. Campisi, A. Neri, "User Adaptive Fuzzy Commitment for Signature Templates Protection and Renewability," *SPIE Journal of Electronic Imaging, Special Issue on Biometrics*, 2008.
16. P. Campisi, E. Maiorana, A. Neri, "On-line signature based authentication: template security issues and countermeasures", in "Biometrics: Theory, Methods, and Applications", N. V. Boulgouris, K.N. Plataniotis, and E. Micheli-Tzanakou, editors, Wiley/IEEE (in print) May 2008.
17. E. Maiorana, P. Campisi, A. Neri, "On-line signature authentication: user adaptive template protection and renewability, *SPIE Defense and Security, Mobile Multimedia/Image Processing, Security, and Applications*, Vol. 6982, Orlando, 2008
18. M. Purser, "Introduction to Error-Correcting Codes", Artech House, Boston, 1995.
19. C. Vielhauer, R. Steinmetz, "Handwriting: Feature Correlation Analysis for Biometric Hashes", *EURASIP Journal on Applied Signal Processing, Special Issue on Biometric*, Vol. 4, pp: 542–558, 2004.
20. M. N. Do, M. Vetterli, "The Finite Ridgelet Transform for Image Representation", *IEEE Transactions on Image Processing*, Vol. 12, No. 1, pp: 16–28, 2003.
21. B. Chen, G. Wornell, "Quantization Index Modulation: a Class of Provably Good Methods for Digital Watermarking and Information embedding", *IEEE Transactions on Information Theory*, Vol. 47, 2001.
22. A. Ross, K. Nandakumar, A.K. Jain, "Handbook of Multibiometrics", Springer, USA, 2006.

Direct attacks using fake images in iris verification

Virginia Ruiz-Albacete, Pedro Tome-Gonzalez, Fernando Alonso-Fernandez,
Javier Galbally, Julian Fierrez, and Javier Ortega-Garcia

Biometric Recognition Group - ATVS
Escuela Politecnica Superior - Universidad Autonoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - Campus de Cantoblanco
28049 Madrid, Spain - <http://atvs.ii.uam.es>
{virginia.ruiz, pedro.tome, fernando.alonso, javier.galbally,
julian.fierrez, javier.ortega}@uam.es

Abstract. In this contribution, the vulnerabilities of iris-based recognition systems to direct attacks are studied. A database of fake iris images has been created from real iris of the BioSec baseline database. Iris images are printed using a commercial printer and then, presented at the iris sensor. We use for our experiments a publicly available iris recognition system. Based on results achieved on different operational scenarios, we show that the system is vulnerable to direct attacks, pointing out the importance of having countermeasures against this type of fraudulent actions.

Key words: Biometrics, iris recognition, direct attacks, fake iris

1 Introduction

The increasing interest on biometrics is related to the number of important applications where a correct assessment of identity is a crucial point. The term *biometrics* refers to automatic recognition of an individual based on anatomical (e.g., fingerprint, face, iris, hand geometry, ear, palmprint) or behavioral characteristics (e.g., signature, gait, keystroke dynamics) [1]. Biometric systems have several advantages over traditional security methods based on something that you know (password, PIN) or something that you have (card, key, etc.). In biometric systems, users do not need to remember passwords or PINs (which can be forgotten) or to carry cards or keys (which can be stolen). Among all biometric techniques, iris recognition has been traditionally regarded as one of the most reliable and accurate biometric identification system available [2]. Additionally, the iris is highly stable over a person's lifetime and lends itself to noninvasive identification because it is an externally visible internal organ [3].

However, in spite of these advantages, biometric systems have some drawbacks [4]: *i*) the lack of secrecy (e.g. everybody knows our face or could get our fingerprints), and *ii*) the fact that a biometric trait can not be replaced (if we forget a password we can easily generate a new one, but no new fingerprint can

be generated if an impostor “steals” it). Moreover, biometric systems are vulnerable to external attacks which could decrease their level of security. In [5] Ratha *et al.* identified and classified eight possible attack points to biometric recognition systems. These vulnerability points, depicted in Figure 1, can broadly be divided into two main groups:

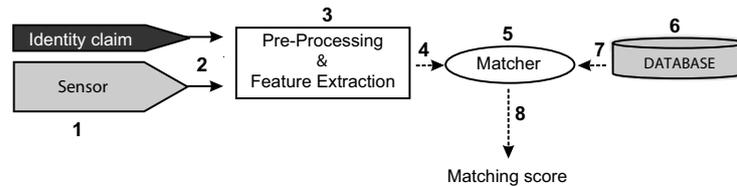


Fig. 1. Architecture of an automated biometric verification system. Possible attack points are numbered from 1 to 8.

- **Direct attacks.** Here, the sensor is attacked using synthetic biometric samples, e.g. gummy fingers (point 1 in Figure 1). It is worth noting that in this type of attacks no specific knowledge about the system is needed. Furthermore, the attack is carried out in the analog domain, outside the digital limits of the system, so digital protection mechanisms (digital signature, watermarking, etc) cannot be used.
- **Indirect attacks.** This group includes all the remaining seven points of attack identified in Figure 1. Attacks 3 and 5 might be carried out using a Trojan Horse that bypasses the system modules. In attack 6, the system database is manipulated. The remaining points of attack (2, 4, 7 and 8) exploit possible weak points in the communication channels of the system. In opposition to direct attacks, in this case the intruder needs to have some additional information about the internal working of the system and, in most cases, physical access to some of the application components. Most of the works reporting indirect attacks use some type of variant of the hill climbing technique introduced in [6].

In this work we concentrate our efforts in studying direct attacks on iris-based verification systems. For this purpose we have built a database with synthetic iris images generated from 27 users of the BioSec multi-modal baseline corpus [7]. This paper is structured as follows. In Sect. 2 we detail the process followed for the creation of the fake iris, and the database used in the experiments is presented. The experimental protocol, some results and further discussion are reported in Sect. 3. Conclusions are finally drawn in Sect. 4.

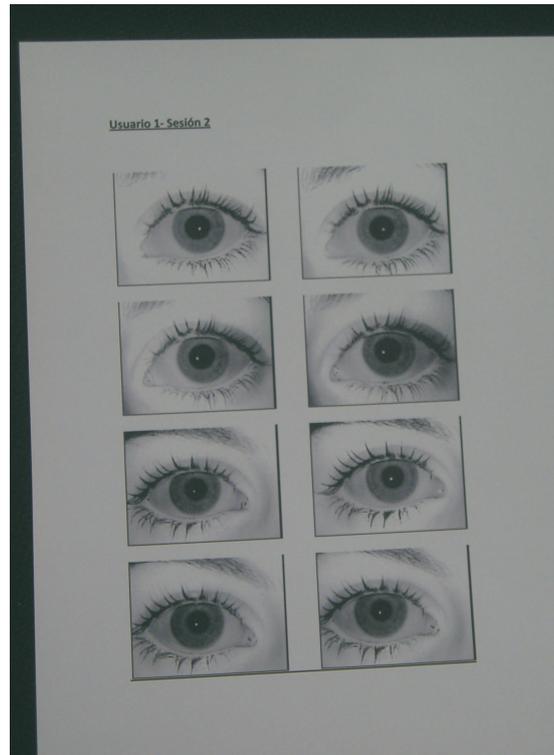


Fig. 2. Iris capture preparation.

2 Fake Iris Database

A new iris database has been created using iris images from 27 users of the BioSec baseline database [7]. The process is divided into three steps: *i*) first original images are preprocessed for a better afterwards quality, then *ii*) they are printed on a piece of paper using a commercial printer as shown in Figure 2, and lastly, *iii*) printed images are presented at the iris sensor, as can be seen in Figure 3, obtaining the fake image.

2.1 Fake iris generation method

To correctly create a new database, it is necessary to take into account factors affecting the quality of acquired fake images. The main variables with significant importance for iris quality are found to be: preprocessing of original images, printer type and paper type.

We tested two different printers: a HP Deskjet 970cxi (inkjet printer) and a HP LaserJet 4200L (laser printer). They both give fairly good quality. On the other hand, we observed that the quality of acquired fake images depends



Fig. 3. Capturing fake iris.

PRINTER	PAPER	PREPROCESSING [8]
Ink Jet Laser	White paper Recycled paper Photographic paper High resolution paper Butter paper Cardboard	Histogram equalization Noise filtering Open/close Top hat

Table 1. Options tested for fake iris generation.

on the type of paper used. Here comes the biggest range of options. All the tested types appear in Table 1. In our experiments, the preprocessing is specially important since it has been observed that the iris camera does not capture correctly original images printed without previous modifications. Therefore we have tested different enhancement methods before printing in order to acquire good quality fake images. The options tested are also summarized in Table 1. By analyzing all the possibilities with a few images, the combination that gives the best segmentation results and therefore the best quality for the afterwards comparison has been found to be the inkjet printer, with high resolution paper and an Open-TopHat preprocessing step. In Figure 4, examples using different preprocessing techniques with this kind of paper and inkjet printer are shown.

2.2 Database

The fake iris database follows the same structure of the original BioSec database. Therefore, data for the experiments consists of $27 \text{ users} \times 2 \text{ eyes} \times 4 \text{ images} \times 2 \text{ sessions} = 432$ fake iris images, and its corresponding real images. Acquisition of fake images has been carried out with the same iris camera used in BioSec, a LG IrisAccess EOU3000.

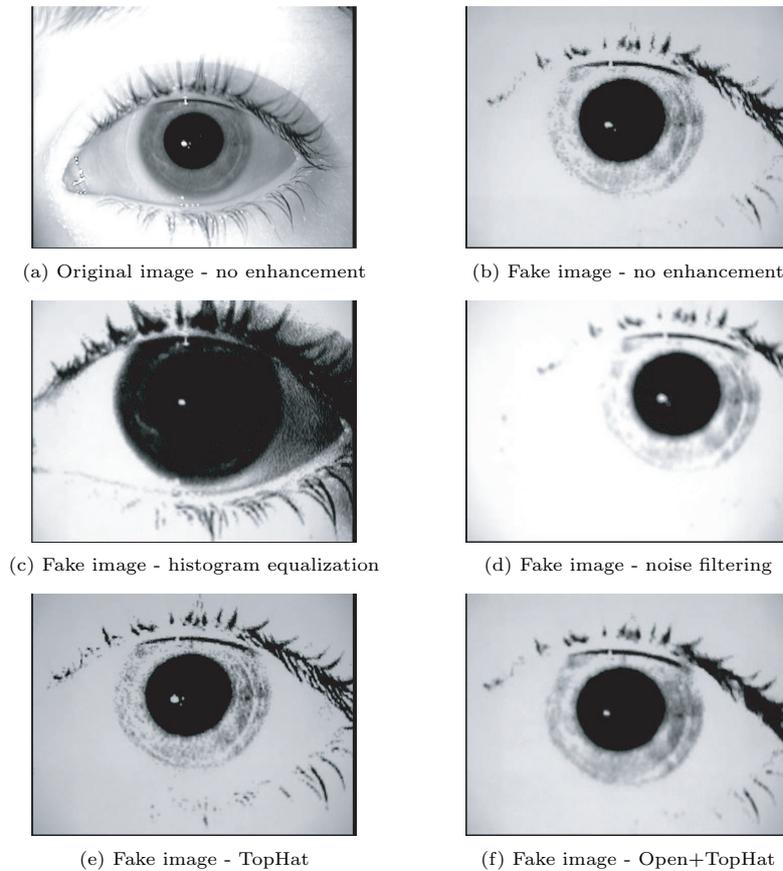


Fig. 4. Acquired fake images with different modifications using high quality paper and inkjet printer.

3 Experiments

3.1 Recognition system

We have used for our experiments the iris recognition system¹ developed by Libor Masek [9]. It consists of the following sequence of steps that are described next: segmentation, normalization, encoding and matching.

For iris segmentation, the system uses a circular Hough transform in order to detect the iris and pupil boundaries. Iris boundaries are modeled as two circles. The system also performs an eyelids removal step. Eyelids are isolated first by fitting a line to the upper and lower eyelid using a linear Hough transform (see

¹ The source code can be freely downloaded from www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html

Figure 5(a) right, in which the eyelid lines correspond to the border of the black blocks). Eyelashes detection by histogram thresholding is available in the source code, but it is not performed in our experiments. Although eyelashes are quite dark compared with the surrounding iris region, other iris areas are equally dark due to the imaging conditions. Therefore, thresholding to isolate eyelashes would also remove important iris regions. However, eyelash occlusion has been found to be not very prominent in our database.

Normalization of iris regions is performed using a technique based on Daugman's rubber sheet model [10]. The center of the pupil is considered as the reference point, based on which a 2D array is generated consisting of an angular-radial mapping of the segmented iris region. In Figure 5, an example of the normalization step is depicted.

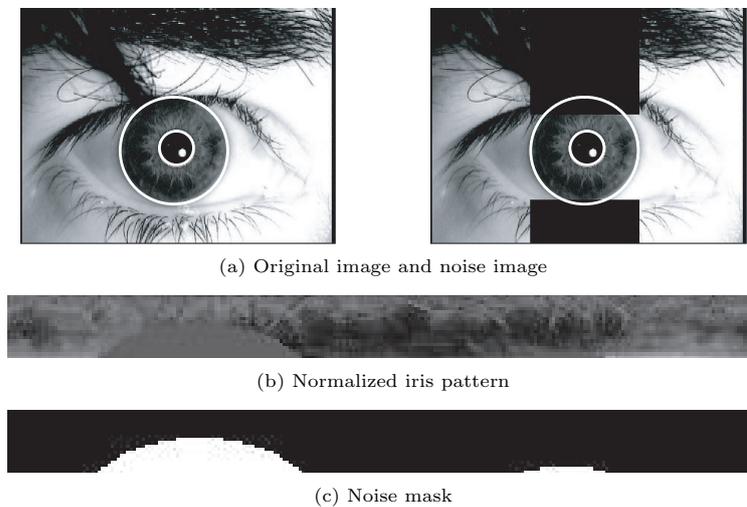


Fig. 5. Examples of the normalization step.

Feature encoding is implemented by convolving the normalized iris pattern with 1D Log-Gabor wavelets. The rows of the 2D normalized pattern are taken as the 1D signal, each row corresponding to a circular ring on the iris region. It uses the angular direction since maximum independence occurs in this direction. The filtered output is then phase quantized to four levels using the Daugman method [10], with each filtering producing two bits of data. The output of phase quantization is a grey code, so that when going from one quadrant to another, only 1 bit changes. This will minimize the number of bits disagreeing, if say two intra-class patterns are slightly misaligned and thus will provide more accurate recognition [9]. The encoding process produces a binary template and a corresponding noise mask which represents the eyelids areas (see Figure 5 (c)).

For matching, the Hamming distance is chosen as a metric for recognition. The Hamming distance employed incorporates the noise mask, so that only significant bits are used in calculating the Hamming distance between two iris templates. The modified Hamming distance formula is given by

$$HD = \frac{1}{N - \sum_{k=1}^N Xn_k(OR)Yn_k} \cdot \sum_{j=1}^N X_j(XOR)Y_j(AND)Xn'_j(AND)Yn'_j$$

where X_j and Y_j are the two bitwise templates to compare, Xn_j and Yn_j are the corresponding noise masks for X_j and Y_j , and N is the number of bits represented by each template.

In order to account for rotational inconsistencies, when the Hamming distance of two templates is calculated, one template is shifted left and right bitwise and a number of Hamming distance values are calculated from successive shifts [10]. This method corrects for misalignments in the normalized iris pattern caused by rotational differences during imaging. From the calculated distance values, the lowest one is taken.

3.2 Experimental Protocol

For the experiments, each eye in the database is considered as a different user. In this way, we have two sessions with 4 images each for 54 users (27 donors \times 2 eyes per donor).

Two different attack scenarios are considered in the experiments and compared to the system normal operation mode:

- **Normal Operation Mode (NOM)**: both the enrollment and the test are carried out with a real iris. This is used as the reference scenario. In this context the FAR (False Acceptance Rate) of the system is defined as the number of times an impostor using his own iris gains access to the system as a genuine user, which can be understood as the robustness of the system against a zero-effort attack. The same way, the FRR (False Rejection Rate) denotes the number of times a genuine user is rejected by the system.
- **Attack 1**: both the enrollment and the test are carried out with a fake iris. In this case the attacker enrolls to the system with the fake iris of a genuine user and then tries to access the application also with a fake iris of the same user. In this scenario an attack is unsuccessful (i.e. the system repels the attack) when the impostor is not able to access the system using the fake iris. Thus, the attack success rate (SR) in this scenario can be computed as: $SR = 1 - FRR$.
- **Attack 2**: the enrollment is performed using a real iris, and tests are carried out with fake iris. In this case the genuine user enrolls with his/her iris and the attacker tries to access the application with the fake iris of the legal user. A successful attack is accomplished when the system confuses a fake iris with its corresponding genuine iris, i.e., $SR = FAR$.

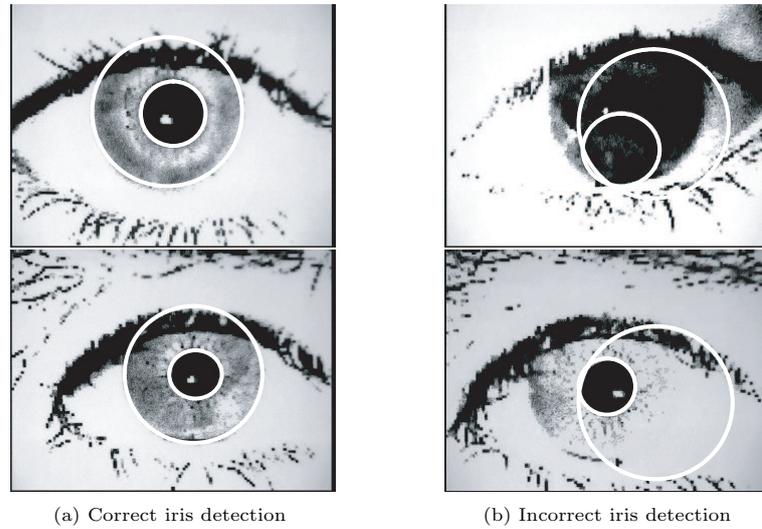


Fig. 6. Examples of fake images with correct iris detection (left) and incorrect iris detection (right).

In order to compute the performance of the system in the normal operation mode, the experimental protocol is as follows. For a given user, all the images of the first session are considered as enrolment templates. Genuine matchings are obtained by comparing the templates to the corresponding images of the second session from the same user. Impostor matchings are obtained by comparing one randomly selected template of a user to a randomly selected iris image of the second session from the remaining users. Similarly, to compute the FRR in attack 1, all the fake images of the first session of each user are compared with the corresponding fake images of the second session. In the attack 2 scenario, only the impostor scores are computed matching all the 4 original samples of each user with its 4 fake samples of the second session. In our experiments, not all the images were segmented successfully by the recognition system. As a result, it was not possible to use all the eye images for testing experiments.

3.3 Results

The number of correctly segmented images were 348 for the original database (80.56% of the 432 available) and 166 for the fake database (38.43% of the 432). In Figure 6, several examples of fake images with correct and incorrect iris detection are plotted. The rate of correctly segmented images for the original database is consistent with that reported in the description of the recognition system used in this paper, with which a segmentation rate of around 83% is attained on the CASIA database [9]. Regarding fake images, it is worth noting that nearly 40% of them pass through the segmentation and normalization stages, and they are

input into the feature extraction and matching stages. It should be noted that the version of the CASIA database used in [9] provided good segmentation, since pupil regions of all iris images were automatically detected and replaced with a circular region of constant intensity to mask out the specular reflections, thus making iris boundaries clearly distinguishable.

In Table 2 we show the Success Rate (SR) of the direct attacks against the recognition system at four different operating points, considering only the matchings between correctly segmented images. The decision threshold is fixed to reach a FAR={0.1, 1, 2, 5} % in the normal operation mode (NOM), and then the success rate of the two proposed attacks is computed. We observe that in all the operating points, the system is highly vulnerable to the two attacks (i.e. a success rate of 50% or higher is observed). This is specially evident as the FAR in the normal operation mode is increased. Also, higher success rates are observed for attack 1. For this kind of attack, an intruder would be correctly enrolled in the system using a fake image of another person and at a later date, he/she would be granted access to the system also using a fake image.

NOM	Attack 1	Attack 2
FAR - FRR (%)	SR (%)	SR (%)
0.1 - 12.71	57.41	49.32
1 - 8.70	74.07	66.06
2 - 7.86	76.85	68.78
5 - 6.19	82.41	73.30

Table 2. Evaluation of the verification system to direct attacks. NOM refers to the system normal operation mode and SR to the success rate of the attack.

4 Conclusion

An evaluation of the vulnerabilities to direct attacks of iris-based verification systems has been presented. The attacks have been evaluated using fake iris images created from real iris of the BioSec baseline database. We printed iris images with a commercial printer and then, we presented the images to the iris sensor. Different factors affecting the quality of acquired fake images have been studied, including preprocessing of original images, printer type and paper type. We have chosen the combination giving the best quality and then, we have built a database of fake images from 54 eyes, with 8 iris images per eye. Acquisition of fake images has been carried out with the same iris camera used in BioSec.

Two attack scenarios have been compared to the normal operation mode of the system using a publicly available iris recognition system. The first attack scenario considers enrolling to the system and accessing it with fake iris. The second one represents accessing a genuine account with fake iris. Results

showed that the system is highly vulnerable to the two evaluated attacks. We also observed that about 40% of the fake images were correctly segmented by the system. When that this happens, the intruder is granted access with high probability, being the success rate of the two attacks of 50% or higher.

Liveness detection procedures are possible countermeasures against direct attacks. For the case of iris recognition systems, light reflections or behavioral features like eye movement, pupil response to a sudden lighting event, etc. have been proposed [11, 12]. This research direction will be the source of future work. We will also explore the use of another type of iris sensors, as the OKI's hand-held iris sensor used in the CASIA database².

Acknowledgments. This work has been supported by Spanish project TEC2006-13141-C03-03, and by European Commission IST-2002-507634 Biosecure NoE. Author F. A.-F. is supported by a FPI Fellowship from Consejería de Educación de la Comunidad de Madrid. Author J. G. is supported by a FPU Fellowship from the Spanish MEC. Author J. F. is supported by a Marie Curie Fellowship from the European Commission.

References

1. Jain, A., Ross, A., Pankanti, S.: Biometrics: A tool for information security. *IEEE Trans. on Information Forensics and Security* **1** (2006) 125–143
2. Jain, A., Bolle, R., Pankanti, S., eds.: *Biometrics - Personal Identification in Networked Society*. Kluwer Academic Publishers (1999)
3. Monro, D., Rakshit, S., Zhang, D.: DCT-Based iris recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **29**(4) (April 2007) 586–595
4. Schneier, B.: The uses and abuses of biometrics. *Communications of the ACM* **48** (1999) 136
5. Ratha, N., Connell, J., Bolle, R.: An analysis of minutiae matching strength. *Proc. International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA Springer LNCS-2091* (2001) 223–228
6. Soutar, C., Gilroy, R., Stoianov, A.: Biometric system performance and security. *Proc IEEE Workshop on Automatic Identification Advanced Technologies, AIAT* (1999)
7. Fierrez, J., Ortega-Garcia, J., Torre-Toledano, D., Gonzalez-Rodriguez, J.: BioSec baseline corpus: A multimodal biometric database. *Pattern Recognition* **40**(4) (April 2007) 1389–1392
8. Gonzalez, R., Woods, R.: *Digital Image Processing*. Addison-Wesley (2002)
9. Masek, L., Kovesi, P.: Matlab source code for a biometric identification system based on iris patterns. The School of Computer Science and Software Engineering, The University of Western Australia (2003)
10. Daugman, J.: How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology* **14** (2004) 21–30
11. Daugman, J.: Anti spoofing liveness detection. available on line at <http://www.cl.cam.ac.uk/users/jgd1000/countermeasures.pdf>
12. Pacut, A., Czajka, A.: Aliveness detection for iris biometrics. *Proc. IEEE Intl. Carnahan Conf. on Security Technology, ICCST* (2006) 122–129

² <http://www.cbsr.ia.ac.cn/databases.htm>

Evaluating systems assessing face-image compliance with ICAO/ISO standards¹

M. Ferrara, A. Franco, D. Maltoni

C.d.L. Scienze dell'Informazione - Università di Bologna, via Sacchi 3, 47023 Cesena, ITALY
DEIS – Viale Risorgimento, 2 – 40126 Bologna, Italy.
E-mail: {ferrara, franco, maltoni}@csr.unibo.it

Abstract. This paper focuses on the requirements for face images to be used in Machine Readable Travel Documents, defined in the ISO/IEC 19794-5 standard. In particular an evaluation framework is proposed for testing software able to automatically verify the compliance of an image to the standard. The results obtained for three commercial software are reported and compared.

1. Introduction

Face represents one of the most used biometric traits, for both computer automated and human assisted person identification. To allow interoperability among systems developed by different vendors and simplify the integration of biometric recognition in large-scale identification (e-passport, visas, etc.) a standard data format for digital face images is needed. In this context, the International Civil Aviation Organization (ICAO) started in 1980 a project focused on machine assisted biometric identity confirmation of persons. Initially three different biometric characteristics were identified for possible application in this context (face, fingerprint, iris), but finally face was selected as the most suited to the practicalities of travel document issuance, with fingerprint and/or iris available for choice by States for inclusion as complementary biometric technologies. Of course high quality, defect-free digital face images are needed to maximize both the human and computer assisted recognition accuracy. Starting from the ICAO work, in 2004 the International Standard Organization (ISO) defined a standard [3] for the digital face images to be used in the Machine Readable Travel Documents. The standard specifies a set of characteristics that the image has to comply, mainly related to the

¹This work was partially supported by CNIPA (Centro Nazionale per l'Informatica nella Pubblica Amministrazione).

position of the face in the image and to the absence of defects (blurring, red eyes, face partially occluded by accessories, etc.) that would affect both the human and automatic recognition performance.

In view of the widespread adoption of the new standard, some vendors of biometric technologies started to develop and distribute software applications able to automatically verify the compliance of a face image to the ISO standard. However, until now no independent and systematic evaluation of these algorithms have been done, and it is not clear if these systems can effectively assist or substitute humans in checking face-image compliance with the standards.

To the best of our knowledge one of the few experiments related to this issue has been carried out by the Federal Office for Information Security (BSI) in Germany, one of the first European countries to adopt the electronic passport; this evaluation [5], aimed at verifying the compliance of face images to the ISO/IEC 19794-5 standard [3], was performed on 3000 images from field applications, and was carried out mainly by manual inspection.

The aim of this paper is to define a testing protocol for the automatic evaluation of systems verifying compliance of face-images with ISO/IEC 19794-5 standard. Starting from the guidelines and the examples of compliant and non-compliant images provided in the ISO standard, a set of salient characteristics has been identified and encoded, a precise evaluation protocol has been defined, and a software framework has been developed to fully automate the test. We believe that the possibility of fully automating such evaluation is a crucial point since it allows to effortlessly repeat the test on new systems and new databases.

The paper is organized as follows: in section 2 the main ISO requirements are detailed, in section 3 the evaluation protocol and framework are introduced; section 4 presents the experiments carried out and finally in section 5 some concluding remarks are given.

2. The ISO/IEC 19794-5 standard and the tests defined

The ISO/IEC 19794-5 international standard [3] specifies a record format for storing, recording and transmitting the facial image information and defines scene constraints, photographic properties and digital image attributes of facial images.

Each requirement is specified for different face image types:

- *Full frontal*. Face image type that specifies frontal images with sufficient resolution for human examination as well as reliable computer face recognition. This type of image includes the full head with all hair in most cases, as well as neck and shoulders.
- *Token frontal*. Face image type that specifies frontal images with a specific geometric size and eye positioning based on the width and height of the image. This image type is suitable for minimizing the storage requirements and to simplify computer based recognition (the eyes are in a fixed position).

The requirements introduced by the ISO standard are organized in two categories: geometric and photographic requirements.

The *geometric requirements* are related to the position of the face and of its main components (eyes, nose, etc.) within the digital image. In Fig. 1. the geometric characteristics of the digital image used to specify the requirements for the full frontal format are shown. The following basic elements are considered in the definition of the requirements:

- A: image width, B: image height;
- AA: imaginary vertical line positioned at the center of the image;
- BB: vertical distance from the bottom edge of the image to an imaginary horizontal line passing through the center of the eyes;
- CC: head width defined as the horizontal distance between the midpoints of two imaginary vertical lines; each imaginary line is drawn between the upper and lower lobes of each ear and shall be positioned where the external ear connects the head;
- DD: head height defined as the vertical distance between the base of the chin and the crown.

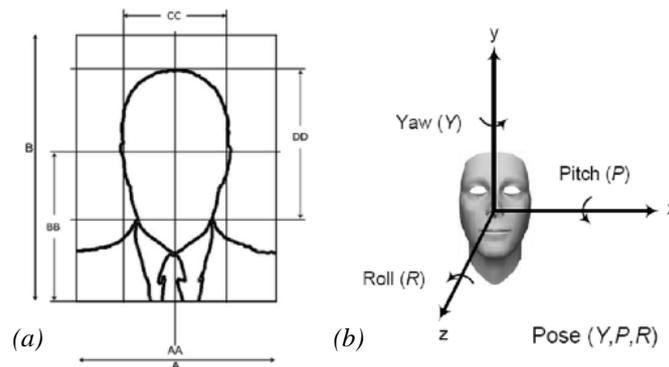


Fig. 1. Geometric characteristics of the Full Frontal Face Image (a) and definition of the pose angles with respect to the frontal view of the subject (b).

The *photographic requirements* refer to characteristics of the face (e.g. expression, mouth open) and of the image (e.g. focus, contrast, natural skin tone). Starting from the guidelines and the examples of acceptable/unacceptable images provided in [3], we defined a set of tests (see Table 1).

Table 1. Tests defined to evaluate systems for ISO compliance check. The last column of the table (Section) denotes the section of [3] from which the test was derived.

N°	Description of the test	Section
Feature extraction accuracy tests		
1	Eye Location Accuracy	
2	Face Location Accuracy (other points)	
Geometric tests (Full Frontal Image Format)		
3	Eye Distance (min 90 pixels)	8.4.1
4	Relative Vertical Position ($0.5B \leq BB \leq 0.7B$)	8.3.3
5	Relative Horizontal Position (no tolerances)	8.3.2
6	Head Image Width Ratio ($0.5A \leq CC \leq 0.71A$)	8.3.4
7	Head Image Height Ratio ($0.7B \leq DD \leq 0.8B$)	8.3.5
Photographic and pose-specific tests		
8	Blurring	7.3.3
9	Looking Away	7.2.3
10	Ink Marked/Creased	A3.2.3
11	Unnatural Skin Tone	7.3.4
12	Too Dark/Light	7.3.2
13	Washed Out	7.4.2.1
14	Pixelation	A3.2.3
15	Hair Across Eyes	A3.2.3
16	Eyes Closed	7.2.3
17	Varied Background	A2.4
18	Roll/Pitch/Yaw Greater 5	7.2.2
19	Flash Reflection on Skin	7.2.10
20	Red Eyes	7.3.4
21	Shadows Behind Head	A3.2.3
22	Shadows Across Face	7.2.7
23	Dark Tinted Lenses	7.2.11
24	Flash Reflection on Lenses	7.2.11
25	Frames too Heavy	A4.3
26	Frame Covering Eyes	7.2.3
27	Hat/Cap	A3.2.3
28	Veil over Face	A3.2.3
29	Mouth Open	7.2.3
30	Presence of Other Faces or Toys too Close to Face	A3.2.3

The token face image format inherits the requirements of the frontal face image type [3], does not require to comply with the geometric constraints of full frontal images (see tests 3..7 in Table 1), but enforces

other geometric constraints related to the eyes position and the image size proportion (see Table 2).

Table 2. Geometric tests for the token image type. The last column of the table (Section) denotes the section of [3] from which the test was derived.

Geometric tests (Token Frontal Image Format)	Section
Image Width W (min 240 pixels)	9.2.4
Image Height ($= W / 0.75$)	9.2.3
Y Coordinate of Eyes ($=0.6 * W$)	9.2.3
X Coordinate of First (right) Eye ($=(0.375 * W) - 1$)	9.2.3
X Coordinate of Second (left) Eye ($=(0.625 * W) - 1$)	9.2.3
Width from Eye to Eye (inclusive) ($=0.25 * W$)	9.2.3

3. The software framework

A software framework has been developed to evaluate and analyze the performance of algorithms provided in the form of SDK (Software Development Kit). The framework offers the following functionalities.

- *Manual image labeling.* It allows to load a database of images and, for each of them, to manually:
 - label by point and click the main facial features such as eye centers, center of mouth, nostrils, etc.;
 - specify the compliance of the image with respect to the characteristics underlying the tests 8.30 in Table 1. Labels are tri-state values (compliant, non-compliant and *dummy*). A *dummy* label is assigned when the human expert is not confident enough whether the image is compliant or not.
- *Artificial dataset generation.* Most of the images used for the tests belong to face databases available to the scientific community; for some of the tests it is very difficult to find in these databases a sufficient number of non-compliant images. The framework offers a tool to generate artificial images non-compliant with respect to a particular characteristic by applying some image processing to “real” compliant images. In the current version of the framework the following transformations are available: blurring, brightness and contrast adjustment, pixelation, addition of red eyes. Each transformation is characterized by a specific set of parameters that can be tuned to control the effect of the operation on the real image (see figure 2 for some example).

- *Automatic SDK testing.* In order to interface the framework with different SDKs a simple interface protocol based on a command-line executable has been defined and provided to the SDK vendors. The executable evaluates the compliance of a single image and provides in output a compliance degree (in the range 0..100) for each of the characteristics underlying the tests in Table 1. The results obtained can be analyzed and compared on the basis of several performance indicators among which EER, FAR/FRR curves, DET graphs. Any new SDK, in order to be tested, simply needs to comply with the defined testing protocol.

4. Experiments

Three commercial SDKs, whose names cannot be disclosed (here referred to as A, B and C), have been evaluated in our experiments; for each of them the compliance of each image in the dataset has been measured with respect to the characteristics 1, 8..30 underlying the cases reported in Table 1; the geometric tests 2..7 are not included in this study, because of the non-uniform way the different SDKs provide in output details about the location of internal face-feature. This part of the evaluation will be done in a successive study.

Analogously to a biometric verification systems the SDKs here evaluated can make two types of errors: declaring compliant with respect to a given characteristic an image that is non-compliant (False Acceptance) and declaring non-compliant and image that is compliant (False Rejection). Images labeled as dummy for a given characteristic are excluded from the corresponding test.

According to this protocol, the results are reported for each characteristic in terms of EER and rejection rate. A rejection occurs when either the SDK is not able to process an image or the image is processed but the SDK is not able to evaluate the specific characteristic. According to the best practices the rejection is here included in the calculation of EER [4]: this is implicitly done by, assuming that a 0 compliance degree (for the given characteristic) is returned in case of rejection. This choice is aimed at discouraging the software to reject the most uncertain cases thus improving the performance over processed images.

4.1 The database

The dataset used is the publicly available AR Face Database [1] containing 2000 images of different subjects. Unfortunately some of the images are defective or not available, so that finally 504 images from 126 subjects have been selected. The images, whose original size is 768×576 pixels, have been cropped to 480 (w) \times 640 (h). The database contains about four images of each subject: one with natural lighting and expression, two with evident facial expressions (smile and angry) and one with a strong lateral illumination. The presence of images with varying expression and lighting allows to verify the ability of the various SDKs to evaluate the compliance with respect to some of the characteristics given in section 2. Unfortunately the original dataset contains no (or a few) images non-compliant with respect to some of the requirements identified. In order to carry out a more precise evaluation of all the characteristics, some additional “artificial” datasets have been generated by applying specific digital image operations (see Fig. 2) that cannot be described here in detail for lack of space. In particular, derived datasets have been generated for blurring (1008 images), unnatural skin tone (748), too dark/light (735), washed out (1008), pixelation (1008), red eyes (1008). The images in these datasets are equally distributed between compliant and non-compliant.

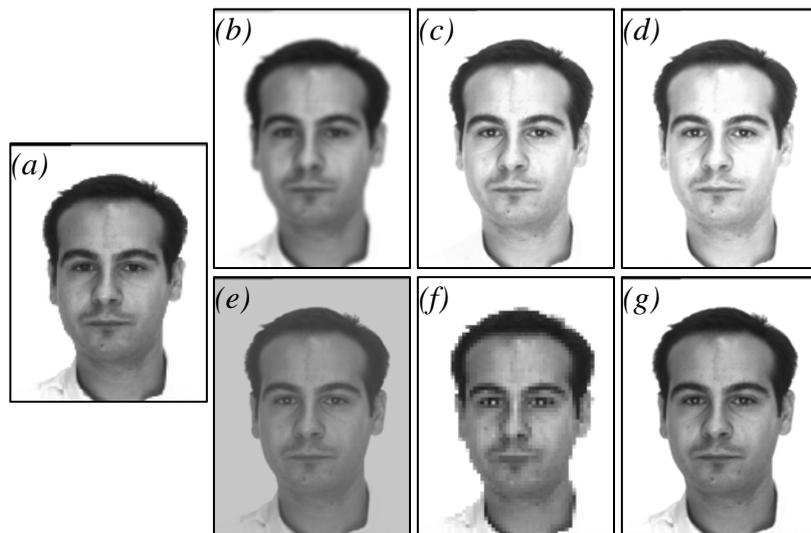


Fig. 2. An example of application of each image operation. (a) Original image; (b) blurred; (c) unnatural skin tone; (d) too dark/light; (e) washed out; (f) pixelation; (g) red eyes.

4.2 Experimental results

The results of the evaluation carried out are reported in this section. As to the geometric requirements, the eye localization accuracy of the SDKs is shown in Table 3. The columns refer to increasing intervals of localization errors (in pixels). The column “Correct” includes all the cases where the maximum error (among the single errors for the two eyes) is lower than 6 pixels. On average, in the images used for testing, the distance between the two eyes is 125 pixels. The result for SDK B is not reported since it does not output the eye position. The two SDKs achieve a very good localization accuracy, even in the presence of difficult cases; A is the most accurate.

Table 3. Eye localization accuracy.

SDK	Correct	[6;9[[9;13[[13;16[[16;∞[
A	488	4	4	3	5
C	424	64	5	4	7

Table 4. EER and Rejection Rate of the three SDKs evaluated.

Characteristic	A		B		C	
	EER	Rej.	EER	Rej.	EER	Rej.
8 Blurred	1.88%	3.47%	2.48%	0.30%	65.87%	0.60%
9 Looking Away	1.79%	0.20%	-	-	1.19%	0.40%
10 Ink Marked/Creased	-	-	-	-	-	-
11 Unnatural Skin Tone	7.09%	0.00%	50.00%	0.13%	2.67%	0.40%
12 Too Dark/Light	-	-	25.15%	0.14%	25.17%	0.54%
13 Washed Out	-	-	23.11%	0.99%	0.79%	1.98%
14 Pixelation	-	-	1.39%	0.50%	-	-
15 Hair Across Eyes	50.00%	94.44%	-	-	-	-
16 Eyes Closed	12.11%	2.90%	-	-	22.59%	0.41%
17 Varied Background	17.91%	0.24%	48.87%	0.72%	46.86%	0.48%
18 Roll/Pitch/Yaw Greater 5	-	-	13.96%	0.60%	43.72%	0.40%
19 Flash Reflection on Skin	0.51%	0.20%	49.38%	0.60%	-	-
20 Red Eyes	4.86%	0.60%	50.00%	0.99%	3.70%	1.10%
21 Shadows Behind Head	-	-	-	-	-	-
22 Shadows Across Face	28.94%	2.78%	-	-	34.77%	0.40%
23 Dark Tinted Lenses	-	-	-	-	25.00%	0.40%
24 Flash Reflection on Lenses	-	-	-	-	22.77%	0.42%
25 Frames too Heavy	-	-	-	-	-	-
26 Frame Covering Eyes	50.00%	93.82%	-	-	16.18%	0.20%
27 Hat/Cap	-	-	-	-	-	-
28 Veil over Face	-	-	-	-	/	0.40%
29 Mouth Open	5.88%	23.72%	-	-	14.64%	0.20%
30 Objects too Close to Face	-	-	-	-	-	-
- the SDK does not support the test for this characteristic / the EER is not calculated since the dataset does not contain non-compliant images The bolded values indicate the best performance for each characteristic. The grayed rows correspond to characteristics evaluated mainly on compliant images. For these characteristics additional tests on extended datasets are needed.						

The results obtained by the three SDKs on tests 8..30 are reported in Table 4 where the EER and rejection rate are given. The rejection rate is in most cases quite low, but it is worth noting that this value for SDK A is noticeable for some characteristics (e.g. hair across eyes). For a further comparison of the three SDKs, the results in terms of EER shown in Table 4 are summarized in Fig. 3 where the EER distribution for the three SDKs is reported. Five EER intervals have been defined and each bar of the graph represents the number of tests that a given SDK is able to manage with an accuracy value included in the related range.

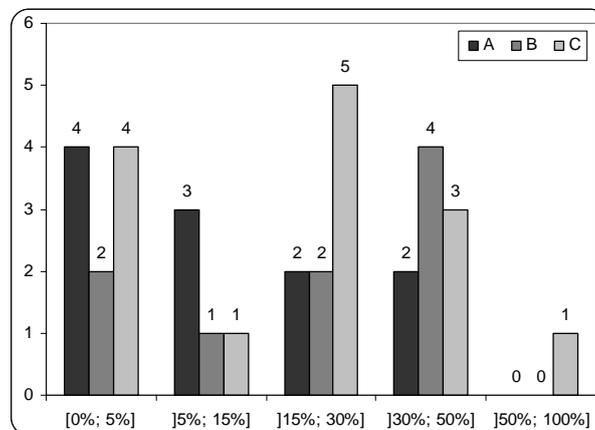


Fig. 3. Distribution of the three SDKs accuracy in five EER intervals. The x-axis reports the EER ranges, and the y-axis indicates the number of tests on which a SDK reaches an EER included in that range.

It is worth noting that the number of characteristics evaluated by the three SDKs is different: in particular, A verifies 11 requirements, reaching in most cases a good accuracy; B evaluates only 9 requirements and the results obtained are mostly unsatisfactory; finally, C deals with 14 requirements and the accuracy is quite variable and strictly dependent on the specific requirement.

5. Conclusions

This work addresses the problem of evaluating the accuracy of automatic software for ISO/IEC 19794-5 compliance check. To this purpose, a testing protocol and an evaluation framework have been developed.

The results show that the three SDKs evaluated are able to accurately

check only some characteristics while achieving unsatisfactory performance for others. An analysis of the results in Table 4 and Fig. 3 show that some requirements (e.g. blurred, unnatural skin tone, washed out) are easily verifiable by an automatic software. On the other hand, characteristics like hair across eyes or frame covering eyes are difficult to be automatically evaluated, and a human expert inspection is recommended. Finally, characteristics such as looking away, too dark/light and mouth open are not classified accurately by the three SDKs, but a deeper analysis of the problem and the availability of training images would certainly allow to significantly improve the performance. As future work the dataset will be extended by including new samples of non-compliant images with respect to all the grayed characteristics in Table 4. It is our intention to make a new database (labeled and partitioned into training and test sets) available to the scientific community to allow the comparison with other SDKs, the improvement of existing techniques and development of new algorithms.

Bibliography

- [1] A.M. Martinez, R. Benavente, "The AR Face Database", *CVC Technical Report #24*, June 1998.
- [2] BSI-TR 03104, "Technical Guideline for production, data acquisition, quality testing and transmission for passports - Annex 1 - Quality requirements for the acquisition and transmission of the facial image data as biometric feature for electronic identification documents", 2007.
- [3] ISO International Standard ISO/IEC JTC 1/SC 37 N506, "Text of FCD 19794-5, Biometric Data Interchange Formats – Part 5: Face Image Data", 2004. Available at <http://isotc.iso.org>
- [4] R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman and A.K. Jain, "Performance Evaluation of Fingerprint Verification Systems", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.28, no.1, pp.3-18, 2006.
- [5] U. Seidel, "Experiences with facial image standard compliance- Do we need to relax ISO 19794-5 tolerances?", *E-Passport Interoperability Test Event*, available at http://www.interoptest-berlin.de/reading_material.htm

Automatic Evaluation of Stroke Slope

Georgi Gluhchev¹, Ognian Boumbarov²

¹Institute of Information Technologies, 2, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
gluhchev@iinf.bas.bg,

² Faculty of Communication Technologies, Technical University, 8, Kl. Ohridski,
1000 Sofia, Bulgaria,
olb@tu-sofia.bg

Abstract. An approach for the evaluation of the average slope of the vertically oriented strokes of signatures is described. It is based on the properties of the Fourier Transform which allow accumulating energy of pixels lying on straight lines of same slope alongside a single straight line in the frequency domain. The slope of the latter determines straightforwardly the stroke slope since both slopes differ in $\pi/2$.

Key words: Handwriting, Line slope, Segmentation

1 Introduction

A new emerging area of interest concerns the increased people mobility and development of fast and reliable authentication systems based on biometric parameters, including handwriting. The ever increasing threats of illegal access to specific information or equipment require developing of reliable and non-abusive access-permit systems. The signature happened to be one of the biometrics modalities that had been commonly accepted and used for document authentication. The identification parameters relate to geometric shape of different elements constituting the signature, type of connections between them, evaluation of the applied pressure and writing dynamics, tilt towards the basic line. To measure them automatically one has to segment first the signature into strokes.

The slope of the strokes is one of the parameters that have been always used. It reflects the established writer's dynamic stereotype. It may play a significant role in cases where no forgery is expected, or in case of specific handwriting. However, strictly determined slope does not exist at all due to the natural variations of handwriting from the one hand, and different stroke slope inherent to specific characters and connections between them, on the other hand. This specificity has lead to the qualitative estimation of the slope in terms of categories as "left", "upright", "right", "predominantly right" and like. But they do not indicate exactly how big the slope is and do not make it possible to distinguish between different "right" slopes for example. Measurement of slope of different strokes is tedious and time consuming work and is prone to subjectivity. By this reason an objective measure of the "average" slope of the strokes in handwriting is desirable.

Another problem related to the graphometric methodology in handwriting analysis concerns the evaluation of character width and distances between letters. If the average slope of writing is known this may help the proper segmentation of the words and contribute to the above mentioned parameters.

Different heuristics and Hough Transform (HT) based techniques have been used for the detection of line slope [1,4,5]. Since HT is generally applied to binarized images, to apply it for halftone images V. Shapiro [4] replaces the original image by a simulated one, using the DH (Digital Halftone) transform and showing the closeness of the obtained results to those obtained from RD (Radon transform) applied to the original half tone image. Thus the computational cost inherent to RD is reduced. However, it is hardly applicable to the problem of stroke slope evaluation. In [1] an attempt in this direction is made, also based on HT, where a method is proposed making the HT-approximation error close to zero. Thus, evaluating maximums in HT-space the row slope and character tilt could be evaluated. The major problem in all HT-based cases concerns the calculation workload.

In this paper an approach is suggested dealing with halftone images and giving the possibility for fast and reliable stroke slope evaluation. It is based on the well known properties of the Fourier Transform (FT) which is an additional advantage, because one can use optimized FT procedures included in the libraries of scientific oriented software products like Matlab.

2 The Approach

The approach is based on the fact stemming from the Fourier slice theorem that FT of a straight line of slope θ is a straight line of slope $\theta + \pi/2$. This statement could be checked in the following way.

Let the image $f(p,q)$ of size $N \times N$ contains only the horizontal line l_o : $q=0$, i.e.

$$f(p,q) = \begin{cases} 1, & \text{if } q=0 \\ 0, & \text{otherwise} \end{cases}.$$

Its DFT (Discrete Fourier Transform) is obtained according to the following formula [2]

$$g(m,n) = \frac{1}{N} \sum_{p=0}^{N-1} \sum_{q=0}^{N-1} f(p,q) \exp(-2\pi j \frac{mp+nq}{N}) = \frac{1}{N} \sum_{p=0}^{N-1} \exp(-2\pi j \frac{mp}{N}) \quad (1)$$

The summands from the last sum are terms of a geometric progression with a quotient $\exp(-2\pi j / N)$, therefore

$$g(m,n) = \frac{1}{N} \frac{(\exp(-2\pi j m / N))^N - 1}{\exp(-2\pi j m / N) - 1} = \begin{cases} 1, & \text{if } m=0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Thus, DFT of l_o is the vertical straight line at $m=0$, which may be taught as a rotation of l_o at 90° about the origin.

Let now the horizontal line l_o be shifted at the position $q_0 \neq 0$. According to the shifting property of FT [2,3] we will have

$$\hat{g}(m,n) = g(m,n) \exp(-2\pi j \frac{nq_0}{N}) \quad (3)$$

$\hat{g}(m,n) = 0$ for $m > 0$ and an arbitrary $n < N$ since $g(m,n) = 0$ in that case. At $m=0$

the value of $\hat{g}(0,n)$ will be $\hat{g}(0,n) = \exp(-2\pi j \frac{nq_0}{N})$, i.e., again the only non-zero

column will be the first one. This result shows that if we have a few horizontal straight line segments in $f(p,q)$ their DFT will result in a non-zero column at the origin. This is so because the image $f(p,q)$ could be presented as a sum of as many as the number of segments images, each of them containing just one segment. Therefore, DFT of a set of horizontal lines will be an image of zero entries except the ones alongside the first vertical column.

In the same way using the rotation property of FT we may claim that the FT of a line l_θ of slope θ will result in the rotation at angle θ of the FT of line l_o . Therefore the FT of the rotated line will be a non-zero line rotated at the angle $\theta + \pi/2$. Same will be valid for a set of line segments oriented at an angle θ . This suggests the following technique for the detection and extraction of straight lines of same slope in a source image $f(p,q)$.

1. Evaluate $g(m,n)$ as a centered DFT of $f(p,q)$.
2. Using a circular scan of $g(m,n)$ about the center detect the peaks alongside the circle.
3. Evaluate the angle of line through the center and the maximal peak.

3 Experimental results

To check the efficacy of the approach different experiments have been carried out in Matlab environment. In the first experiments gray scale images with line segments of almost same orientation were used. In Fig. 1 slanted line segments are present together with their DFT. An angle of 76.2° is detected from the image in Fig. 2b which corresponds to -13.8° of slope for the original image.

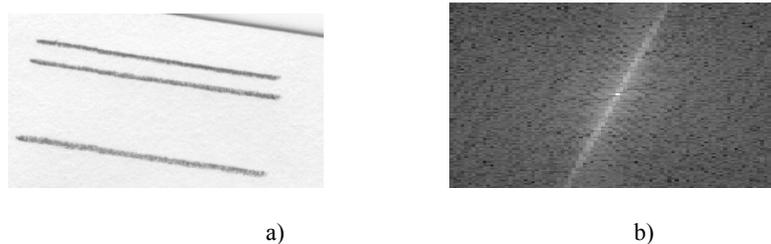


Fig. 1. a) Slant lines, b) DFT of the image from a)

The background around the bright line in Fig 2b is not uniform which is due to the non-zero background in the original image. Also, the lines drawn by pencil are not fully black and of same width as it could be seen in Fig. 1a. This may cause problems with the correct detection of the slope. To avoid random bright pixels in the DFT image that may produce false maximums, it is better to use different radii of circular scanning. In these simple cases a value of about $1/6$ of the image height was used without problems.

The above examples contained simple images. To be useful in practice the approach has to have the ability to detect the slope in more complicated cases, consisting of strokes of different orientation, provided predominant orientation exists, e.g. signatures.

In Fig. 2a) an example of a signature is shown. Fig. 2b presents its DFT and Fig. 2c shows the plot of the circular scan. An angle of 75° was evaluated corresponding to the highest maximum. This maximum is related to the vertically oriented strokes. The other two bright lines describe the slope of the upper signature strokes inclined less than $\pi/4$ and the slope of the intermediate connecting elements.

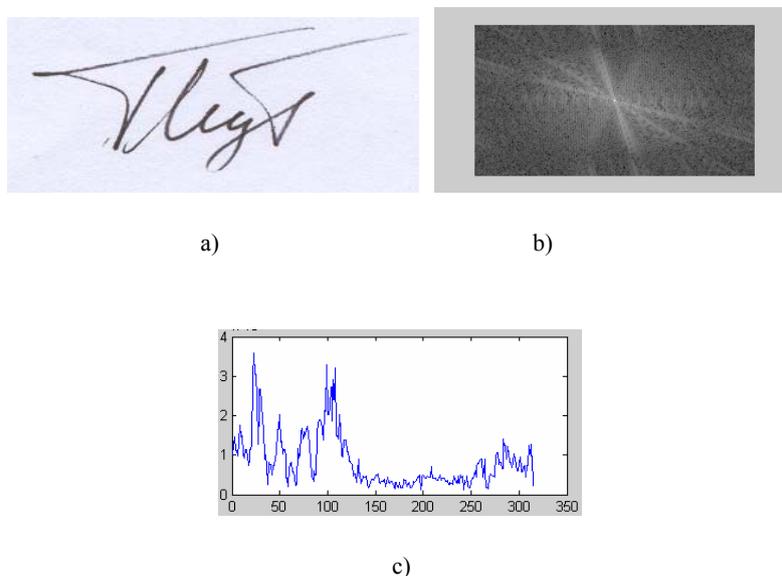


Fig. 2. a) Signature, b) DFT of the signature, c) Plot of the circular scan

Detecting the slope angle one could apply inverse FT preserving only the values alongside the corresponding line in the FT domain. Fig. 3a) presents the result of such an operation for the image in Fig. 2a). Applying a proper threshold one will obtain the image in Fig. 3b) where the lines correspond to the major vertically oriented strokes in the signature. Distances between them may be used as another quantitative identification parameter. This could be easily achieved if a projection on the line perpendicular to the strokes is generated. The distances between the local peaks will correspond to the differences between strokes.



Fig. 3. a) Inverse FT alongside the slope, b) Extraction of the black lines

One could repeat this operation using the next peak from the circular scanning graph thus obtaining the major horizontally oriented strokes.

3 Conclusion

In this paper a DFT (Discrete Fourier Transform) based approach is applied to the automatic detection of the slope of straight line segments in signatures. It does not assume a binarized image as an input. Basic properties of FT are used to prove its adequacy. Using the well developed procedures for the evaluation of FT, the approach does not require many efforts for its implementation and it is computationally inexpensive. The preliminary experiments have shown that it is robust to the background structure and to the quality of the foreground object. The extraction of predominantly oriented strokes could help signature segmentation and evaluation of important authentication parameters.

Acknowledgments. This investigation was supported by the Ministry of Education and Science in Bulgaria, contract No BY–TH-202/2006.

References

1. Dimov, D. Using an Exact Performance of Hough transform for Image Text Segmentation, Proc. of IEEE Conf. on Image Processing ICIP'01, vol.1, Greece, 2001, pp. 778-781
2. Petrou M., P. Bosdogianni, *Image Processing*, John Wiley & Sons, Inc., 1999
3. Pratt, W.K. "Digital Image Processing" (third Edition), John Wiley & Sons, Inc., 2001
4. Shapiro, V. "On the Hough Transform of Multilevel Pictures", *Pattern Recognition*, 29(4), 1996, pp. 589-602
5. Likforman-Sulem, L., A. Hanimyan, C. Faure. "A Hough Bazed Algorithm for Extracting Text Lines in Handwritten documents", Proc. of Int. Conf. Document Analysis and Recognition, vol. II, Canada, 1995, pp. 774-777

Biometric system based on voice recognition using multiclassifiers

Mohamed CHENAF^(1,2), Dan ISTRATE⁽¹⁾, Valeriu VRABIE⁽²⁾, and Michel
HERBIN⁽²⁾

⁽¹⁾ESIGETEL, 1 Rue du Port de Valvins, 77210 Avon-Fontainebleau, France

⁽²⁾CRéSTIC, Université de Reims, Moulin de la Housse, 51687 Reims, France

{mohamed.chenafa,dan.istrate}@esigetel.fr

{valeriu.vrabie,michel.herbin}@univ-reims.fr

Abstract. In this paper we present a new speaker recognition system based on the fusion of two identification classifiers followed by a verification step. The user pronounces two passwords: the first one is composed by three words uniquely combined from a set of 21 possible words, while the second password represents the name of the user. The first step of the proposed system uses the first password to feed two identification classifiers: a speaker identification system (text independent) and a isolated word identification system (speaker independent). The isolated word identification system is constructed as the fusion of three classifiers, one for each word of the first password. The aim of this first step is to identify a couple speaker/password corresponding to the first password by combining the results of the two identification classifiers. A verification system is then applied on the second password in order to confirm or infirm the identification result (speaker identity) given by the fusion above. Compared with a state of the art speaker recognition system (text dependent) that gives an EER of 4.76%, the first step of the proposed system provides an EER of 0.38%, while the second step an EER of 0.26% for a text independent verification and of 0.13% for a text dependent verification.

Key words: Biometric recognition system, Speaker identification, Isolated word recognition, Data fusion, GMM/UBM.

1 Introduction

The biometric recognition systems, used to identify persons on the basis of physical or behavioral characteristics (voice, fingerprints, face, iris, etc.), have gained in popularity during recent years especially in forensic work and law enforcement applications [1]. The use of the voice as a biometric characteristic offers advantages such as: it is well accepted by users, can be recorded by regular microphones, the hardware costs are reduced, etc. Two different tasks can be defined for voice-based biometric systems: speaker identification and speaker verification. In the former case, an unknown speaker is compared to N known speakers models stored in the database and the best matching speaker is returned as the

recognition decision. In the later case, an identity is claimed by a speaker and the system compares the voice sample to the voice model of the claimed speaker. If the similarity exceeds a predefined threshold, the speaker is accepted, otherwise is rejected. Two methods can be employed for both systems: text-dependent and text-independent. The text pronounced by the speaker is known beforehand by the system in the former case, while the system does not have any information on the pronounced text in the later case [9].

However, due to channel distortions, ambient noise, etc., a mismatch between training and testing conditions appears and the performances of voice-based biometric systems easily degrade. In order to improve the performances of these systems a solution is to merge different information carried out by the speech signal. Several studies on data fusion shown that the performances of this kind of speaker recognition system are improved [2, 8, 10]. However, the results are less good compared to biometric systems based on other modalities (fingerprint, iris, etc) or on the fusion of different modalities.

This paper proposes a fusion approach that uses two kinds of information contained in the speech signal: the speaker (who spoke?) and the password pronounced (what was said?). A first test signal is used to identify a couple speaker/password. This step is done by merging the likelihood ratios given by two identification systems: a speaker identification system (text-independent) and a word recognition system (speaker-independent). The word identification system is constructed as the fusion of three isolated word recognition systems, one for each word of the first test signal. The speaker identified by this first step is then confirmed by a classical verification system on a second test signal. The first test signal is composed by three words uniquely combined from a set of 21 possible words, while the second one represents the name of the user. The proposed system gives good improvements in terms of Equal Error Rate (EER) compared with the state of the art (text dependent speaker recognition system). Note that the experiments presented in this study use the platform ALIZE developed by the LIA laboratory (Avignon University, France) [4].

This paper is organized as follows. Section 2 provides a brief overview of speaker recognition systems. Section 3 presents the proposed system while the experiments are discussed in Section 4, followed by conclusions in the last section.

2 Speaker recognition system overview

The general structure of an automatic speaker recognition system is shown in Figure 1. This system operates in two modes: training and recognition. In the training mode a new speaker (with a known identity) is enrolled into the database, while in the recognition mode an unknown speaker gives a speech input signal and the system try to identify the speaker. This system can be used for both identification and verification tasks.

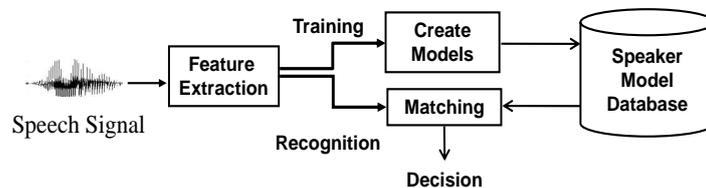


Fig. 1. Architecture of a speaker recognition system.

2.1 Features extraction

Features extraction is the first component of an Automatic Speaker Recognition (ASR) system [6]. It transforms the input speech waveform into a sequence of acoustic feature vectors (called also parameters) through a signal time division. Most of the speech parameters used in speaker recognition systems relies on a cepstral representation of the speech signal [11]. The aim of this transformation is to obtain a new representation that is more compact, less redundant, and more suitable for statistical modeling.

Mel-frequency cepstral coefficients (MFCC): The use of the MFCC parameters is motivated by studies of the human peripheral auditory system. The speech signal $x(n)$ is firstly divided into q short time windows. The Discrete Fourier Transform (DFT) is then applied to convert each time window into the spectral domain. Each magnitude spectrum is then smoothed by a bank of triangular overlapping bandpass filters. Each filter, $H(k, m)$, computes a weighted average of that sub-band, which is then logarithmically compressed:

$$X'(m) = \ln \left(\sum_{k=0}^{N-1} |X(k)| H(k, m) \right), \quad (1)$$

where $X(k)$ is the DFT of a time window of length N of the signal $x(n)$, the index k corresponds to the frequency $f_k = kf_s/N$, with f_s the sampling frequency, the index m is the filter number, and the filters $H(k, m)$ are triangular filters defined by the center frequencies $f_c(m)$ [13]. The log compressed filter outputs $X'(m)$, called also *mel log-amplitudes*, are then decorrelated by using the Discrete Cosine Transform. The MFCCs are the amplitudes of the resulting spectrum.

A schematic representation of this procedure is given in Figure 2.

The mel mapping used here to define the bank of triangular filters is:

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right). \quad (2)$$

The LFCC parameters are calculated in the same way as the MFCC, but the triangular filters use a linear frequency repartition.

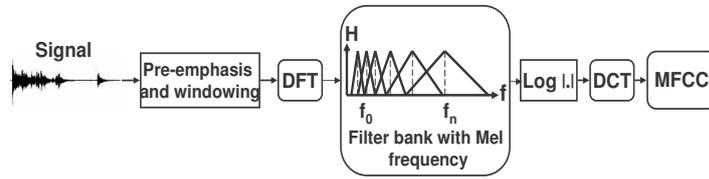


Fig. 2. Extraction of MFCC parameters.

Energy and Derivatives(Δ , $\Delta\Delta$): Usually we need to add other parameters to the cepstral ones, such as the energy and the derivatives. The energy in a frame is the sum over time of the power of the samples in the frame. Another important fact about the speech signal is that it is not constant from one frame to another, for this reason we also add features related to the change in cepstral features over time. We do this by adding for each vector features a velocity feature (Δ) and acceleration feature ($\Delta\Delta$)[10].

2.2 Speaker modeling

The training phase uses the acoustic vectors extracted from each segment of the signal to create a speaker model which will be stored in a database. In ASR system there are two class of methods that give good results of recognition: deterministic methods (dynamic comparison and vector quantization) and statistical methods (Gaussian Mixture Model - GMM, Hidden Markov Model - HMM), these last ones being the most used in this domain.

We have chosen to use a GMM based system that employs a Universal Background Model (UBM). The UBM has been introduced by [12] in speaker verification in order to capture the general characteristics of a population. This model is created by using all recording of the database, the aim being to have a general model of speakers which will be then used to adapt each speaker model. This choice was motivated by two reasons: modeling by GMM is very flexible with regard to the type of the signal and the use of GMM gives a good compromise between performances and the complexity of the system.

GMM-UBM: The matching function in GMM is defined in terms of the log likelihood of the GMM in respect to the speech segment X given by:

$$p(X|\lambda) = \sum_{q=1}^Q \log p(x_q|\lambda), \tag{3}$$

where $p(x_q|\lambda)$ is the Gaussian mixture density of the q^{th} segment in respect to the speaker λ :

$$p(x_q|\lambda) = \sum_{i=1}^G p_i f(x_q|\mu_i^{(\lambda)}, \Sigma_i), \tag{4}$$

with the mixing weights constrained by $\sum_{i=1}^G p_i = 1$.

In these expressions x_q is the D-dimensional acoustic vector corresponding to the q^{th} time window of the input signal, p_i , $\mu_i^{(\lambda)}$ and Σ_i ($i = 1, \dots, G$) are the mixture weight, mean vector, and covariance matrix of the i^{th} Gaussian density function (denoted by f) of the speaker λ , while G denotes the number of GMM used by the model.

The speaker model λ is thus given by: $\lambda = \left\{ p_i, \mu_i^{(\lambda)}, \Sigma_i | i = 1, \dots, G \right\}$. The UBM model has the same form: $UBM = \left\{ p_i, \mu_i^{(UBM)}, \Sigma_i | i = 1, \dots, G_U \right\}$ and is created by using all recordings of the database.

The mean vectors of speaker model $\mu_i^{(\lambda)}$ are adapted to the training data of the given speaker from the UBM, i.e. $\mu_i^{(UBM)}$, by using the Maximum a Posteriori (MAP) adaptation method [7], the covariance matrices and mixture weights remaining unchanged.

2.3 Pattern matching and decision

Given a segment of speech, Y , and a hypothesized speaker, S , the task of speaker recognition system is to determine if Y was spoken by S . This task can be defined as a basic hypothesis test between:

- H_0 : Y is from the hypothesized speaker S
- H_1 : Y is not from the hypothesized speaker S

To decide between these hypotheses, the optimum test is the likelihood ratio:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta & \text{Accept } H_0 \\ < \theta & \text{Reject } H_0 \end{cases}, \quad (5)$$

where $p(Y|H_i)$ is the probability density function for the hypothesis H_i evaluated for the observed speech segment Y , also referred to the likelihood of the hypothesis H_i . The decision threshold for accepting or rejecting H_0 is θ . A good technique to compute the two likelihoods, $p(Y|H_0)$ and $p(Y|H_1)$, is given in [5].

3 Proposed system architecture

In this paper we present a new ASR system based on the fusion of two identification classifiers followed by a verification step (Fig. 3). This system is divided into two stages, the first one composed by two classifiers (speaker and word classifiers) and the second one made up by a verification system using the decision result of the first stage. Each speaker is identified by two signals: the first one (combination of three words from a set of 21 possible words) is used by both speaker and word identification systems, while the second one by the verification system. All classifiers used a normalization UBM model, as presented in section

2.2. This means that during the creation of the models (speakers, words), each model is adapted by the MAP method from the UBM model.

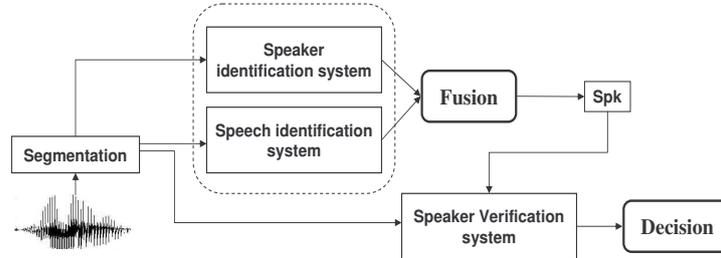


Fig. 3. Global system architecture.

3.1 Speaker identification text independent system

The speaker identification system is an open-set text independent system. This system calculate the log likelihood ratio, by using Eq. (3), between the first signal (made up by three words W_1, W_2, W_3) and all speakers models. No decision is taken at this level, but the *log* likelihood ratios are sorted.

3.2 Word identification speaker independent system

The same signal, made up by three words, is also used to feed a word identification speaker independent system (Fig. 4). This system is constructed as the fusion of three classifiers, one for each word of the first signal. The outputs of each classifier are used in order to propose one or several recognized combinations of words. Only the first three outputs of each module are combined by taking into account the log likelihoods and the validity of the password. Each combination of outputs will have associated the sum of their log likelihood. This approach, which uses a manual words segmentation, was compared with a Viterbi algorithm that performs an automatic extraction of the three words from the entire first signal. The results are presented in section 4.

3.3 Data Fusion

After sorting the *log* likelihood ratios $LLK(W_1, W_2, W_3 | Sp_i)$ calculated with regard to the models of each speaker Sp_i , with $i = 1 \dots N$ and N the number of speakers stored in the database, and the *log* likelihood ratios $LLK(W_1, W_2, W_3 | Psw_i)$ calculated with regard to the models of each password Psw_i (see Fig. 4), a first test consists to compare the most likely speaker given by the speaker classifier with the first three identified passwords (made up by three words) given by the word identification system. If his password was found between the three identified passwords, a couple (speaker/password) was thus identified. A

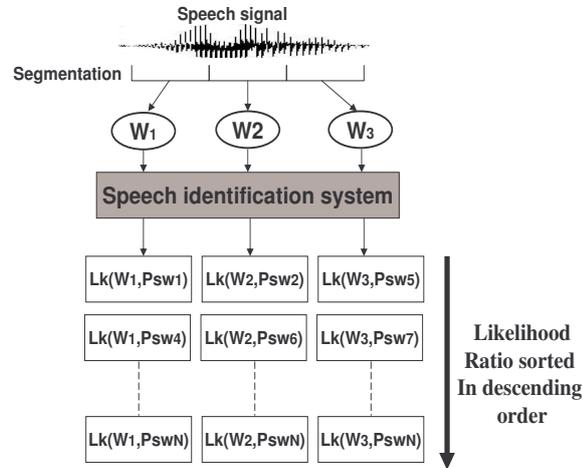


Fig. 4. Speech identification system.

second test consists to compare the most likely password with the first three identified speakers. If this password belongs to one of them, another couple (password/speaker) is identified. In the cases where two couples are identified, the couple with the biggest likelihood ratio ($Lk_Sp + Lk_Psw$) is retained. The system can reject directly a recording if there are no identified couples.

3.4 Speaker verification system

The verification system uses a second signal pronounced by the speaker previously identified in section 3.3. If the likelihood ratio of this verification is smaller than a predefined threshold, the identity of the speaker is confirmed, otherwise the speaker is rejected. For this stage, we have tested two possibilities: a verification based on a text-independent system (no information available on the pronounced word) and on a text-dependent system. The former is more flexible as it allows the speaker to pronounce what he prefers.

4 Experiments

Database: In order to evaluate the proposed system a corpus of specific passwords has been recorded. This corpus contains the recordings of 21 isolated words (French language) pronounced by 5 woman and 53 man ($\approx 4, 28$ hours). The recordings were stored in WAV format, with a sampling rate $f_s = 44.1kHz$.

Parameterization : The parameterization was realized by using MFCC parameters for the passwords modeling and LFCC for the speaker modeling. We

have optimized the acoustic parameter for this application; all the 8 ms the signal is characterized by a vector made up of 16 cepstrals coefficients, energy and their derivatives (Δ , $\Delta\Delta$).

Universal Background Model (UBM) In our experiments, we have tested different sizes (number of Gaussian component, i.e. G_U) of UBM: 64, 128, and 512. Note that the UBM is created by using all recordings of the database. The best compromise performance-computation time was obtained by using $G_U = 128$ Gaussian components for UBM model.

Reference system The results obtained by the global proposed system are compared to a classical text-dependent identification speaker system [3]. In the training stage of the reference system a speaker model is created from the feature vectors (16 LFCC+Energy+ Δ + $\Delta\Delta$). Each speaker model is created by using three passwords (made up by three words). However the recognition phase uses all the passwords of the speakers pronounced by the impostors and the other two passwords pronounced by the clients. We have optimized the number of Gaussian components for the tests signal. The optimal value for our database is $G = 24$ best components from the $G_U = 128$ of UBM model.

Training and test data: For both identification systems (speaker and password) the first signal was composed by three words W_1, W_2, W_3 combined in a unique way from the set of 21 possible words. This kind of signal was used for both training and test steps. The second signal used for the verification system (for training and test also) was chosen in our application as the name of the user. However, this is not a restriction for a text-independent verification system.

The database was divided into two groups: 49 clients and 9 impostors. In order to evaluate the proposed system we have chosen an equal number of positive and negative tests.

1. The speaker identification system (text-independent) uses 3 recordings of 17 words of the 49 clients for the training phase (≈ 29 minutes). For the recognition phase, the system uses 2 recording of 20 words of the 49 clients and 5 recordings of 20 words of the 9 impostors (784 tests).
2. The word identification system (speaker-independent) uses 3 recordings of the 49 clients for the training phase (≈ 29 minutes). For the recognition phase, the system uses 2 recording of the 49 clients and all recordings of the impostors (784 tests).
3. The verification system uses 8 recordings of the second passwords of every client for the training phase (≈ 7 minutes) and 2 recordings of the 49 clients as well as all the recordings of the 9 impostors for the recognition phase.

The reference system uses for the training phase 3 recordings of 3 words of the 49 clients (≈ 8 minutes). For the recognition phase we used 2 recording of the 49 clients and 3 recordings of the 9 impostors (576 tests).

4.1 Results and discussion

Table 1 presents the performances of each stage of the proposed system compared to the state of the art system in terms of Equal Error Rate (EER).

Table 1. performances of different systems.

Systems	Parameters	EER (%)	
Reference System text dependent	16 LFCC+Energy+ $\Delta\Delta$	4.76%	
Fusion between speaker and word identification systems	Isolated words: 16 MFCC+Energy+ $\Delta\Delta$ Speaker: 16 LFCC+Energy+ $\Delta\Delta$	0.38%	
Verification after fusion	16 LFCC+Energy+ $\Delta\Delta$	Dep. 0.13%	Indep. 0.26%

The first stage of the proposed system (fusion) has an EER of 0.38% in comparison with the state of the art (text dependent speaker recognition system) which reach 4.76%. The combination of the speaker identity with the password recognition improves the performances by 90%. The second stage of the system (verification stage) improves the results with 31% in respect to the first stage using a text-independent (EER of 0.26%) and with 34% using text-dependent verification system (EER of 0.13%). In the text independent case, the user has more flexibility: he needs to memorize only a password (of three words) and he can use any text for the second part. The better performance was obtained for a text dependent verification stage, which is explained by the fact that the model contains speaker and text information.

In the first stage we use a word identification system, based on the fusion of three words recognition modules, which gives an EER of 5.56%. The segmentation was ideal (manually applied) but we have evaluated also an automatic segmentation system. We have compared this system with a Viterbi algorithm. In the case of Viterbi algorithm, the passwords are modeled by an HMM with three states and the identification system is feed directly with the combined password of three words (without segmentation). The Viterbi approach gives an EER of 23,84%, which is much higher than the fusion of three words recognition systems based on a manual segmentation. This can be explaining by the fact that in HMM modeling is difficult to reject the impostors.

5 Conclusion and perspectives

In this paper, we have presented several experiments to improve the performances of a voice-based biometric system by using two classifiers and a verification system. The fusion of the results of a speaker identification system and a words identification system constitutes the first stage of the proposed system. This stage improves the EER by 90% in comparison with a state of the art

text-dependent system. The second stage is a speaker verification system that uses the result (speaker identified) of the first stage. The aim here is to confirm or infirm the result returned by the fusion system. This second stage allows to reduce the number of impostors accepted by the first stage and improves the results of the fusion by decreasing the EER from 0.38% to 0.13% (in a text dependent system). The global system improves significantly the performances in term of EER with regard to the reference system. Further works should evaluate the impact of an automatic segmentation module and the influence of different additive noises.

References

1. Atkins, W : A testing time for face recognition technology. *Biometric Technology Today*, Vol 147, pp. 195–197 (2001)
2. BenZeghiba, MF., and Boulard, H. User-customized password speaker verification using multiple reference and background models. *Speech Communication*, Vol 8, pp. 1200–1213 (2006)
3. Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Chagnolleau, I., Meignier, S., Merlin, T., Garciya, J., Delacrtaz, D., and Reynolds, D. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, vol.4, pp. 430–451 (2004)
4. Bonastre, J.-F., Wils, F., and Meignier, S. Alize, a free toolkit for speaker recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 737–740 (2005)
5. Doddington, G. Speaker recognition identifying people by their voices. In *Proc. of the IEEE*, vol 73, pp 1651–1664 (1985)
6. Furui, S. Recent advances in speaker recognition. In *Proc. of the First International Conference on Audio and Video-Based Biometric Person Authentication*, vol 1206, pp. 237–252 (1997)
7. Gauvain, J.-L. and Lee, C.-H. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. In *IEEE Trans. on Speech and Audio*, vol 2, pp. 291–298 (1994)
8. Higgins, J. E., Damper, R. I., and Harris, C. J. Information fusion for subband-hmm speaker recognition. In *International Joint Conference on Neural Networks*, vol 2, pp. 1504–1509
9. Kinnunen, T. *Spectral Features for Automatic Text-Independent Speaker Recognition*, PhDThesis, University of Joensuu, Finland (2003)
10. Kinnunen, T., Hautamki, V., and Franti, P. Fusion of spectral feature sets for accurate speaker identification. In *9th International Conference Speech and Computer (SPECOM)*, pp. 361–365 (2004).
11. Lee, C. H., Soong, F., and Paliwal, K. *Automatic Speech and Speaker Recognition*. Springer, London, UK, (1996)
12. Reynolds, D. *Speaker identification and verification using gaussian mixture speaker models*. Elsevier Science Publishers. *Speech Communication*, vol 17, pp. 91–108 (1995)
13. Sigurdsson, S., Petersen, K. B., and Lehn-Schiøler, T. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, pp. 286–28 (2006)

POLYBIO: Multimodal Biometric Data Acquisition Platform and Security System

Anastasis Kounoudes¹, Nicolas Tsapatsoulis², Zenonas Theodosiou¹, and Marios Milis¹

¹ SignalGeneriX Ltd, Arch.Leontiou A' Maximos Court B', 3rd floor,
P.O.Box 51341, 3504, Limassol, Cyprus
{tasos, z.theodosiou, milis}@signalgenerix.com

²Cyprus University of Technology, Arch.Kyprianos Kyprianos, P.O.Box
50329, 3603, Limassol, Cyprus
nicolas.tsapatsoulis@cut.ac.cy

Abstract. Biometrics is the automated method of recognizing a person based on a physiological or behavioural characteristic. Biometric technologies are becoming the foundation of an extensive array of highly secure identification and personal verification solutions. In the last few years there is increasing evidence that technologies based on multimodal biometrics can provide better identification results if proper fusion schemes are accommodated. In this work, we present a novel platform for multimodal biometric acquisition which combines voice, video, fingerprint and palm photo acquisition through an integrated device, and the preliminary fusion experiments on combining the acquired biometrics modalities. The results are encouraging and show clear improvement both in terms of False Acceptance Rate and False Rejection Rates compared to the corresponding single modality approaches. In the current report, fusion was accommodated at the output of the single modalities; however, fusion experimentation is ongoing and further fusion methodologies are under investigation.

Keywords: Biometric fusion, Data Acquisition, GUI, Matlab

1 Introduction

The emergence of automatic identification of an individual by using certain physiological or behavioral traits, has addressed the

problems that plague traditional verification methods such as passwords and ID cards [1]. Biometric authentication requires comparing a registered or enrolled biometric sample. During enrolment a sample of the biometric trait is captured, processed by a computer, and stored for later comparison. A biometric system based on a single biometric identifier for a personal identification is often not able to meet the desired performance requirements. The performance is largely affected by noise in sensed data, non-universality, upper bound on identification accuracy, and spoof attacks [2].

Some of the limitations of a biometric system can be addressed by using a consolidation of multiple sources of biometric information [3,4,5]. A multimodal biometric system combines a variety of biometric identifies in making a personal identification and takes the advantage of the capabilities of each individual biometric. Based on the nature of biometric modalities, multibiometric systems can be classified into six categories including multi-sensor, multi-algorithm, multi-instance, multi-sample, multimodal and hybrid [6].

Multibiometric systems provide a variety of advantages against traditional biometric systems and are able to encounter the performance requirements of various applications [7]. The problem of non-universality is addressed, since sufficient population coverage can be ensured by a multiple traits. Furthermore, multibiometric systems can facilitate the indexing of large-scale databases, can address the problem of noisy data and provide anti-spoofing measures by making it difficult for an impostor to spoof multiple biometric traits of a legitimate enroll individual.

In this paper we present a new multimodal biometric data acquisition platform and security system. The proposed system uses fingerprint, face, voice and palm geometry features of an individual for verification purposes. The paper is organized as follows: Section 2 presents the single modality biometrics for voice fingerprint and hand geometry. Section 3 describes the Biometrics Fusion. The system is detailed in section 4 whereas Section 5 presents the evaluation of the results and related discussion. Finally, conclusions and further work are stated in Section 6.

2 Single modality biometrics

Multibiometric systems use multiple biometric modalities. A brief description of biometrics that used for our system is given below.

2.1 Voice Biometrics / Extraction method

Voice is the natural means of communication for human beings thus making it the most convenient to use biometric. In addition, voice needs inexpensive equipment for capturing and can be deployed in a variety of telephone-based or internet-based applications where other biometrics are impossible to be deployed. Voice biometric is utilised in this work in the form of text-dependent Speaker Verification using concatenated phoneme Hidden Markov Models (HMMs) [8]. The experimental setup included the evaluation of the Speaker Verification performance using the traditional Mel Frequency Cepstral Coefficients (MFCC) [9, 10] while future experiments will involve the Perceptual Linear Prediction (PLP) coefficients [11].

The procedure is initiated when the user is text-prompted a series of utterances by the system in order to capture the speech samples. This procedure is repeated both in the data capture phase where the multimodal biometric database is created, and the verification phase where the captured speech of a specific user is verified against his HMM models or Voiceprint. A front-end feature extractor is incorporated to calculate the voice features, which are used for both the enrolment and the speaker verification phase. In the enrolment phase, speaker-specific phoneme models are created for each reference speaker. In the speaker verification phase, the phoneme concatenation model corresponding to the prompted single-digit sequence is constructed, and the accumulated likelihood of the input speech frames for the model is compared with a threshold to decide whether to accept or reject the speaker. In the case of successful speaker verification, the features of the speech signal are stored for updating the HMM models of the specific speaker. The approach is based on a simple vocabulary consisting of a single digit numbers spoken continuously in sequences such as “2-3-5-7-9”. The advantage is that by training HMM models for the phonemes needed to construct all the single-digits of the vocabulary, the method can employ random sequences for authentication, and thus its robustness to impostors is increased.

2.2 Fingerprint Biometrics / Extraction Method

Fingerprints are probably the more extensively studied biometric. Uniqueness, permanence, easy acquisition and the small size of the acquisition devices (at least the electronic ones) make fingerprints one of the most popular person identification methods. Usage of fingerprints in verification systems is not so common because fingerprint acquisition has been related, for years, with criminal prosecution and, therefore, it raises user annoyance. This prepossession is getting lower, however, mainly due to the extensive usage of fingerprints for user authentication in popular computing systems such as laptops.

Characteristic fingerprint features are generally categorized into three levels [12]: patterns, points and shape. Patterns are the global details of the fingerprint such as ridge flow and pattern type. Although they are not unique, patterns are useful for fingerprint

classification into generic categories such as whorl, left loop, right loop, etc. Points refer to the characteristics or minutiae proposed by Galton [13] and include ridge bifurcations and endings. They have sufficient discriminating power to establish the individuality of fingerprints. Finally, shape features include all dimensional attributes of the ridge such as ridge path deviation, width, pores, edge contour, incipient ridges, breaks, creases, scars, and other permanent details. It is claimed that shape features are permanent, immutable, and unique according to the forensic experts, and if properly utilized, can provide discriminatory information for human identification.

In the context of the proposed multibiometric system we do not enter into a sophisticated feature extraction process for the fingerprint biometric. Instead we have tried to combine level 1 (patterns) and level 3 (shape) features through a smart combination of fractal scanning of image points and frequency analysis of these points. The proposed fingerprint feature extraction method is simple through powerful: A signature S (1D vector, see also Fig. 1) is created for each 2D fingerprint image by using the well known Hilbert fractal [14] (see Fig. 2) which is one of the most popular space filling curves. Then the power spectrum $P_D(S)$ of the signature is computed over a set of frequency bands (see Fig. 3). The vector of power spectrum values in the various frequency bands is used as feature vector for the fingerprint image.

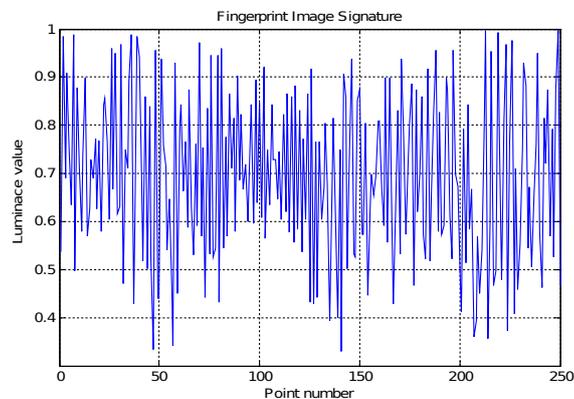


Fig. 1: Image signature using the luminance at sampled points



Fig. 2: Hilbert filling curve for 2D points sampling

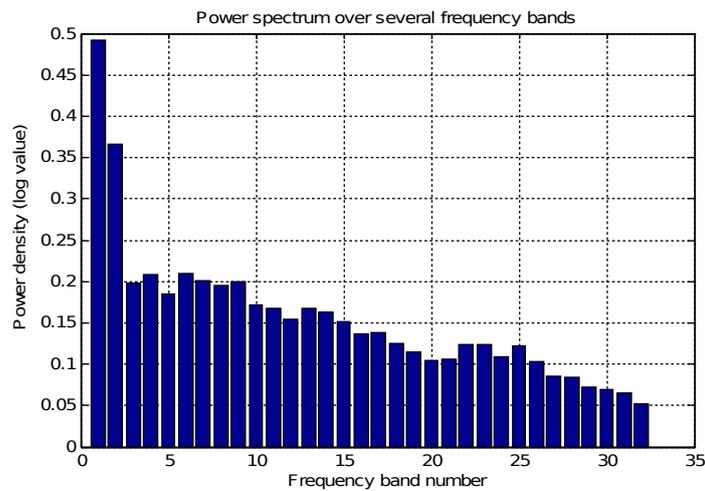


Fig. 3: Feature vector for a fingerprint image

2.3 Hand Geometry Biometrics / Extraction Method

Hand geometry biometric systems are becoming very popular for verification purposes. Although hand geometry is not as unique as other biometrics (e.g., fingerprints), it is permanent and has not been related for criminal prosecution; therefore it is an acceptable method for verification for the great public. In person identification

systems hand geometry has been used mostly as a complement to fingerprints. However, in cases of small user population, hand geometry biometrics are commonly used for authentication since they present acceptable FAR and FRR rates. Hand geometry biometrics fall into two main categories: geometric measurements and contour description. The automatic extraction of geometric measurements from a hand geometry image is a rather difficult error pruned task. The method is more appropriate in a semi automatic environment where a human user indicates the prominent points in the hand contour. Contour description methods have in general lower accuracy but they are more robust in automatic authentication processes.

In this study we have adopted a contour description approach because it is faster and fits well in our multibiometric environment. Fourier descriptors [15] provide a means to describe contours. The idea is to represent the contour as a function of one variable, expand the function in terms of its Fourier series, and use the coefficients of the series as the features.

Let us assume that the palm boundary coordinates $(x(n), y(n))$, $n = 0, 1, \dots, N$, have been extracted in the preprocessing stage. A complex sequence $z(n)$ is simply generated from the boundary coordinates:

$$z(n) = x(n) + jy(n), \quad n = 0, 1, \dots, N - 1 \tag{2.3.1}$$

Taking the Discrete Fourier Transform of the sequence $z(n)$ we get:

$$a(k) = \sum_{n=0}^{N-1} z(n) \exp\left(\frac{-j2\pi kn}{N}\right), \quad 0 \leq k \leq N - 1 \tag{2.3.2}$$

$$\mathbf{a} = [a(0) a(1) \dots a(N - 1)]^T \tag{2.3.3}$$

The values $F_d(k) = \frac{a(k)}{\|\mathbf{a}\|}$ are called Fourier descriptors (please note that there are several types of Fourier descriptors; all of are based on the previously stated principle). It can be easily shown that the values $F_d(k)$ are independent of translation, rotation and scaling.

In the current work we use a limited subset of the Fourier descriptors as the palm geometry biometric:

$$\hat{\mathbf{a}} = [F_d(1) F_d(2) \dots F_d(M)]^T \quad M \ll N \tag{2.3.4}$$

It appears that an M equal to 64 provides an accurate description of the palm contour which is free of noise (see Fig. 4)

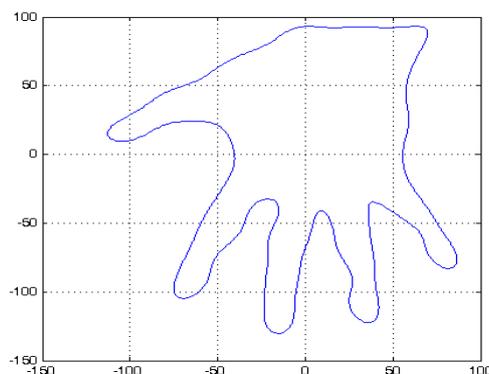


Fig. 4: Palm contour approximation using 64 Fourier coefficients.

3 Biometrics Fusion

An effective fusion scheme is required to combine the information presented by individual modalities. Biometric fusion combines biometric characteristics and can improve accuracy, robustness, fault tolerance and efficiency of a multibiometric system. Three levels of fusion are possible: (a) fusion at the feature extraction, (b) fusion at the matching score level and (c) fusion at the decision level

In the case of fusion at the feature extraction the features obtained from each biometric is used to compute a multimodal feature vector which is used for the biometric authentication. The second approach involves fusion at the matching score level. For each biometric, the user is validated and a matching score indicating the proximity of the feature vector with the trained model is calculated. These scores are then combined in order to verify the claimed identity. The third approach which was used in this work is the fusion at the decision or output level. The final decision is the fusion of individual accept or reject decisions taken by each biometric method.

4 Multibiometric Data Acquisition

Acquiring multimodal biometric data can be a tedious and time consuming task. The use of an integrated system which can provide data collection for a range of different biometrics can greatly simplify the process. For this reason, we have developed POLYBIO [16], a novel, automated system for multimodal biometric data

acquisition. The systems consist of two components: a) The Multimodal biometric sensor hardware shown in Figure 5(a), and b) The Data Acquisition software shown in Figure 5(b). The multimodal biometric sensor hardware integrates an array microphone for voice recording, a digital USB web-camera for face still image and video capture, a USB digital web-camera facing down accompanied by two lighting units and six positioning pins on a black board for palm geometry and a USB optical Fingerprint Reader [17] for fingerprint capture. The hardware component is connected to a PC via a six port USB hub.

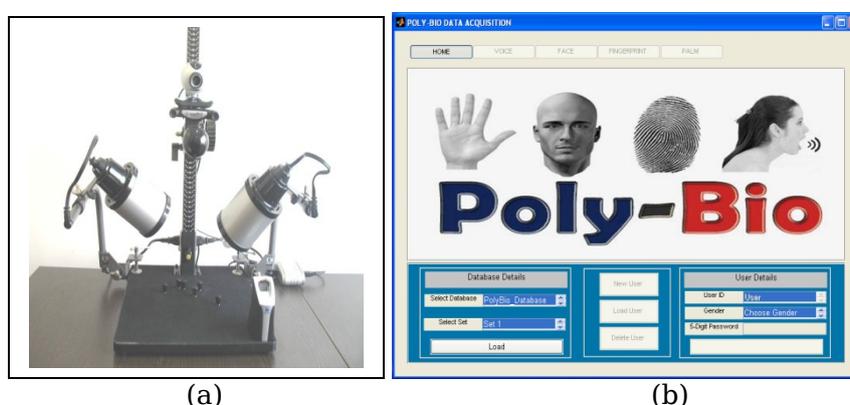


Fig. 5: (a) Multibiometric sensor hardware, **(b)** Data acquisition software

The Data Acquisition software provides a user-friendly Graphical User Interface and an automatic mechanism for capturing and storing data in a multimodal biometric database. The software entails four interactive screens for voice, face, palm, fingerprint data acquisition as illustrated in Figure 6. The administrator can insert a new, select or delete an existing user using the administrator console (Fig. 5(b)). During acquisition, a new entry is created in the system database which contains subfolders for voice, face, palm and fingerprint storage.

A multimodal biometric database was created which contains samples from voice, face, palm and fingerprint for 30 individuals, 15 men and 15 women. Five data capture sessions were stored for each biometric, four of which are used for training and one for testing. The database is used for testing the four biometric methods and for devising data fusion models for improving the overall verification performance.

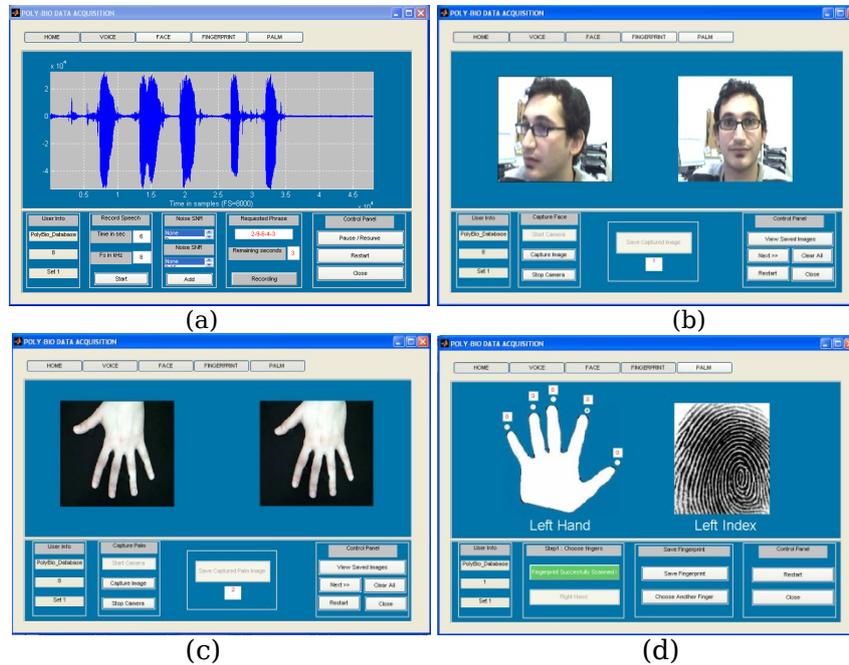


Fig.6: Multimodal Biometric Data acquisition Software screens for **(a)** Voice, **(b)** Face, **(c)** Palm, **(d)** Fingerprint

5 Experimental Results

In this section we present experimental results for biometric authentication based on single modalities (voice, fingerprint, palm geometry) and through fusion of the output scores. As mentioned earlier a multibiometric set of 30 individual was created with three instances per subject used for template creation and the other one for test. In the following paragraphs we describe the verification process in detail.

5.1 Voice verification

Speaker verification performance of the system was evaluated using the MFCC coefficients. Experiments were conducted to assess the effect that the number of the utterances used for training speaker-specific HMM models have on the speaker authentication performance. Tests were also performed to examine the authentication decision threshold selection process and the normalization of HMM scores through the use of a world model.

Single Gaussian mixture HMM models [18] were trained with 13 coefficient MFCC features which include delta coefficients for each speaker using the four enrolment sessions of the database while speaker verification performance was evaluated using 10 utterances from each of the 20 speakers. Each speaker is authenticated against all 20 HMM speaker models using all authentication utterances. The graph in Figure 7(a) was created by averaging speaker dependent HMM scores for each speaker. Axis X shows the speakers attacking each model (impostors) while axis Y shows the speaker dependent HMM models. Axis Z represents the averaged HMM scores for each impostor-model combination. Shifting a horizontal plane along the Z axis and each time taking the point of intersection with Z axis, we calculate the False Acceptance Rate (FAR), False Rejection Rate (FRR) and hence the Equal Error Rate (EER) [11]. In Figure 7(b), the horizontal plane represents the threshold for which FAR equals FRR for the specific experiment. It can be seen that the prominent diagonal represents speaker identification for the 20 speakers.

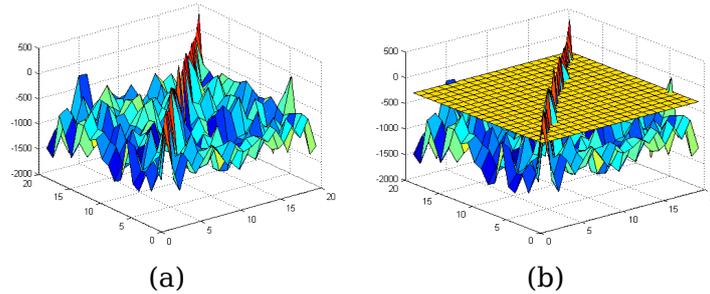


Fig.7: Averaged Speaker Verification Results

Table 1 summarises the evaluation results. It can be seen that better performance was achieved using four enrolment sessions for training and a world model. Even though the best achieved EER=1.8% is not considered adequate for a commercial system, at this stage of the project is acceptable since more research will be performed utilising models with more Gaussian Mixtures, the incorporation of acceleration coefficients, bootstrapping in the training of the models, individual decision threshold for each speaker and Cepstral Mean Subtraction. It is expected that this research will result a significant drop in the EER.

Table 1: Speaker Verification Results

Without World model	Enrolment Sessions		
	2	3	4
% EER	4.12	3.52	2.84

% FAR	4.12	3.53	2.82
% FRR	4.11	3.51	3.86
With World Model	Enrolment Sessions		
	2	3	4
% EER	3.01	2.5	1.8
% FAR	3.00	2.5	1.82
%FRR	3.05	2.5	1.79

5.2 Fingerprint verification

Let us denote with $\mathbf{f}_j^{(k)}$ the j -th fingerprint feature vector of the k -th subject. We have already mentioned that in our experiments we have a population of $N = 30$ subjects (that is $k = 1, 2, \dots, N$) and we use three instances ($j=1, \dots, 3$) per biometric per subject. The fingerprint feature vectors $\mathbf{f}_j^{(k)}$ are the power density values in several frequency bands, described earlier in Section 2.2. We also denote with $\mathbf{y}^{(k)}$ the feature vector used for testing.

Due to the limited number of training instances per subject (i.e., three) we consider as the biometric template of the k -th subject the matrix:

$$\mathbf{F}^{(k)} = [\mathbf{f}_1^{(k)} \ \mathbf{f}_2^{(k)} \ \mathbf{f}_3^{(k)}] \quad (5.2.1)$$

It is obvious that many different templates can be constructed depending on the number of training vectors. Gaussian models and Neural Network representations are among the most popular approaches for template construction and user modeling. In our case we have implicitly consider that all training instances serve as Support Vectors [19].

For each subject we also define a threshold:

$$T^{(k)} = \max_{i \neq j} \left(\left\| \mathbf{f}_i^{(k)} - \mathbf{f}_j^{(k)} \right\| \right) \quad (5.2.2)$$

False Rejection (FR) and False Acceptance (FA) are then defined as:

$$FR : \min_j \left(\left\| \mathbf{y}^{(k)} - \mathbf{f}_j^{(k)} \right\| \right) > T^{(k)} \quad (5.2.3)$$

$$FA : \min_{j, l \neq k} \left(\left\| \mathbf{y}^{(l)} - \mathbf{f}_j^{(k)} \right\| \right) < T^{(k)} \quad (5.2.4)$$

We evaluate the fingerprint biometric by using a four folder cross validation approach. Three instances per subject were randomly selected and used as training patterns while the fourth was used for testing. We repeated this process for 20 cycles and the average

results are shown in the Table 2 below (we report also the experimentation on the number features used).

The limited number of training vectors leads to important FAR and FRR fluctuations. This is mainly due to the adoption of a user specific threshold (see equation (5.2.2)). Including an outlier feature vector in the training set increases the threshold leading to a loose model for the particular subject. This, in turn, increases the FAR for this subject model and may also decrease the FRR. The availability of additional training vectors will alleviate this problem since a more robust threshold would be selected (i.e., based on first order statistics).

Table 2: Average FRR and FAR as a function of feature number for the fingerprint biometric

	Number of features				
	8	12	16	20	32
Average False Rejection Rate (%)	11.2 (\pm 4.5)	9.5 (\pm 3.3)	9.4 (\pm 2.6)	9.1 (\pm 2.3)	8.9 (\pm 2.0)
Average False Acceptance Rate (%)	14.3 (\pm 3.5)	12.6 (\pm 3.2)	10.1 (\pm 2.6)	9.4 (\pm 2.5)	9.0 (\pm 2.6)

5.3 Hand geometry verification

The approach followed for hand geometry verification is identical to the fingerprint verification one. The feature vectors now correspond to the Fourier Descriptors as already mentioned in Section 2.3.

Table 3: Average FRR and FAR as a function of feature number for the hand geometry biometric

	Number of features				
	8	12	16	32	64
Average False Rejection Rate (%)	18.7 (\pm 4.4)	15.7 (\pm 4.1)	12.1 (\pm 2.8)	11.4 (\pm 2.6)	10.7 (\pm 2.5)
Average False Acceptance Rate (%)	16.0 (\pm 4.1)	15.5 (\pm 2.5)	14.8 (\pm 2.4)	9.9 (\pm 2.3)	9.9 (\pm 2.3)

Comparing the results of Tables 2 and 3 it is verified once again the claim that fingerprints are more discriminative than hand geometry. However, the difference is not high; this may be assigned to the simplified feature extraction method adopted for fingerprints.

5.4. Multimodal verification

Our main claim in this work is that multimodal verification can achieve high performance in terms of both FAR and FRR even in cases where single modality verification is not tuned for best performance. This claim is supported by the theory of weak classifiers combination [20] which led to powerful classifiers and pattern recognition systems [21].

We combine the single modalities at the output level using a simple voting scheme: A user is authenticated if the majority of individual modalities vote for authentication and is rejected if the majority vote against.

Table 4 presents the FAR and FRR of the multimodal scheme. In the experimentation we used feature vectors of $M = 20$ elements for the fingerprint biometric and $M = 32$ elements for the hand geometry biometric. The voice print template used is the one obtained via two enrolment sessions and without the usage of World model.

Table 4: Comparison of single modalities and multimodal verification

	Modality			
	Voice	Hand geometry	Fingerprint	Multimodal
Average False Rejection Rate (%)	4.11	11.4	9.1	0.86
False Acceptance Rate (%)	4.12	9.9	9.4	1.23

The results indicate clearly the validity of multimodal verification. The best of single modality FAR and FRR (voice biometric) are far away from the corresponding rates achieved via output level fusion. Furthermore, even the best tuned modality (voice biometric with four enrollment sessions and using world model) does not achieve (FAR = 1.79, FRR = 1.82) the rates obtained by multimodal verification.

6 Conclusions and further work

This study presents an integrated platform for multimodal biometric acquisition for person identification. While the primary aim was to introduce the overall systems we have also presented the methods we use for biometrics extraction from voice, fingerprints and palm contour. It was shown through an experimental study that even weak single modality verification systems can lead to high performance ones using simple fusion methods.

The work on biometric fusion is ongoing. We are currently experimenting on alternative fusion methods including feature-based, score-based and rule-based fusion. In addition we will explore alternative feature extraction methods, at least for the fingerprint and hand geometry modalities. We seek to investigate what happens in cases where highly-tuned single verification modalities are combined through output voting schemes.

Acknowledgments. This work was undertaken in the framework of the POLYBIO (Multibiometric Security System) project funded by the Cyprus Research Promotion Foundation (CRPF) under the contract PLHRO /0506/04.

References

1. Ross, A., Jain, A. K.: Identification Information fusion in biometrics. *Pattern Recognition Letters* 24, 2115–2125 (2003)
2. Ross, A.: An Introduction to multibiometrics. In: *Proc. Of the 15th European Signal Processing Conferene (EUSIPCO)*, Poznan. Poland (2007)
3. Brunelli, R., Falavigna, D.: Person Identification Using Multiple Cues. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol.12, No.10, pp. 955-966 (1995)
4. Bigun, E.S, Bigun, J., Duc, B., Fischer, S.: Expert Conciliation for Multimodal Person Authentication Systems Using Bayesian Statistics. In: *Proc. International Conference on Audio and Video-based Biometric Person Authentication (AVBRA)*, pp. 291-300, Crans-Montana, Switzerland (1997)
5. Kittler, J., Hatef, M., Duin, P.W.R., Matas, J.: On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, no. 3, pp. 226-239 (1998)
6. Ross, A, Nandakumar, K., Jain, A.K.: *Handbook of Multibiometrics*. Springer, New York, USA, 1st edition (2006)
7. Jain, K.A., Ross, A.: *Multibiometric Systems*. *Communications of the ACM*, Special Issue on Multimodal Interfaces (2004)

8. Matsui, T., Furui, S.: Speaker Recognition Using Concatenated Phoneme HMMs. In: Proc. International Conference on Spoken Language Processing, Banff, Th.s AM.4.3 (1992)
9. Mammone, R. J., Zhang, X., Ramachandran, R. P.: Robust Speaker Recognition, A Feature-Based Approach. IEEE Signal Processing Magazine, 13 (5), pp.55-71 (1996)
10. Campbell, J. P.:Speaker Recognition: A Tutorial. In Proc. of the IEEE, 85(9), pp.1437-1462 (1997)
11. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, vol. 87, no. 4, 1738 – 1752 (1990)
12. Pankanti, S., Prabhakar, S., Jain, A.K.: On the Individuality of Fingerprints. IEEE Trans. Pattern Analysis and Machine Intelligence 24(8), 1010-1025 (2002).
13. Galton, F.: Personal Identification and Description. Nature 38, 201-202 (1888).
14. Barra, M., Collado, C., Mateu, J., O'Callaghan, J. M.: Miniaturization of Superconducting Filters Using Hilbert Fractal Curves. IEEE Transactions on Applied Superconductivity 15(3), 3841-3846 (2005).
15. Chellappa, R., Bagdazian, R.: Fourier Coding of Image Boundaries. IEEE Trans. on Pattern Analysis and Machine Intelligence 6(1), 102-105 (1984).
16. POLYBIO website, <http://polybio.signalgenerix.com/>
17. Fingerprint Reader manufacturer, <http://www.bioenabletech.com>
18. Rabiner, L., Juang, B. H.: Fundamentals of Speech Recognition. Prentice Hall (1993)
19. Burges, C. J. C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121 - 167 (1998).
20. Ji, C., Ma, S.: Combinations of weak classifiers. IEEE Trans. on Neural Networks 8(1), 32-42 (1999)
21. Viola, P. Jones, M.: Rapid object detection using a boosted cascade of simple features. Proceeding of CVPR01, vol. 1, 511-518 (2001)

RECENT RESEARCH REPORTS

- #120 Peter Danholt. *Interacting Bodies: Posthuman Enactments of the Problem of Diabetes Relating Science, Technology and Society-studies, User-Centered Design and Diabetes Practices*. PhD thesis, Roskilde, Denmark, February 2008.
- #119 Alexandre Alapetite. *On speech recognition during anaesthesia*. PhD thesis, Roskilde, Denmark, November 2007.
- #118 Paolo Bouquet, editor. *CONTEXT'07 Doctoral Consortium Proceedings*, Roskilde, Denmark, October 2007.
- #117 Kim S. Henriksen. *A Logic Programming Based Approach to Applying Abstract Interpretation to Embedded Software*. PhD thesis, Roskilde, Denmark, October 2007.
- #116 Marco Baroni, Alessandro Lenci, and Magnus Sahlgren, editors. *Proceedings of the 2007 Workshop on Contextual Information in Semantic Space Models: Beyond Words and Documents*, Roskilde, Denmark, August 2007.
- #115 Paolo Bouquet, Jérôme Euzenat, Chiara Ghidini, Deborah L. McGuinness, Valeria de Paiva, Luciano Serafini, Pavel Shvaiko, and Holger Wache, editors. *Proceedings of the 2007 workshop on Contexts and Ontologies Representation and Reasoning (C&O:RR-2007)*, Roskilde, Denmark, August 2007.
- #114 Bich-Liên Doan, Joemon Jose, and Massimo Melucci, editors. *Proceedings of the 2nd International Workshop on Context-Based Information Retrieval*, Roskilde, Denmark, August 2007.
- #113 Henning Christiansen and Jørgen Villadsen, editors. *Proceedings of the 4th International Workshop on Constraints and Language Processing (CSLP 2007)*, Roskilde, Denmark, August 2007.
- #112 Anders Kofod-Petersen, Jörg Cassens, David B. Leake, and Stefan Schulz, editors. *Proceedings of the 4th International Workshop on Modeling and Reasoning in Context (MRC 2007) with Special Session on the Role of Contextualization in Human Tasks (CHUT)*, Roskilde, Denmark, August 2007.
- #111 Ioannis Hatzilygeroudis, Alvaro Ortigosa, and Maria D. Rodriguez-Moreno, editors. *Proceedings of the 2007 workshop on REpresentation models and Techniques for Improving e-Learning: Bringing Context into the Web-based Education (ReTleL'07)*, Roskilde, Denmark, August 2007.
- #110 Markus Rohde. *Integrated Organization and Technology Development (OTD) and the Impact of Socio-Cultural Concepts — A CSCW Perspective*. PhD thesis, Roskilde University, Roskilde, Denmark, 2007.