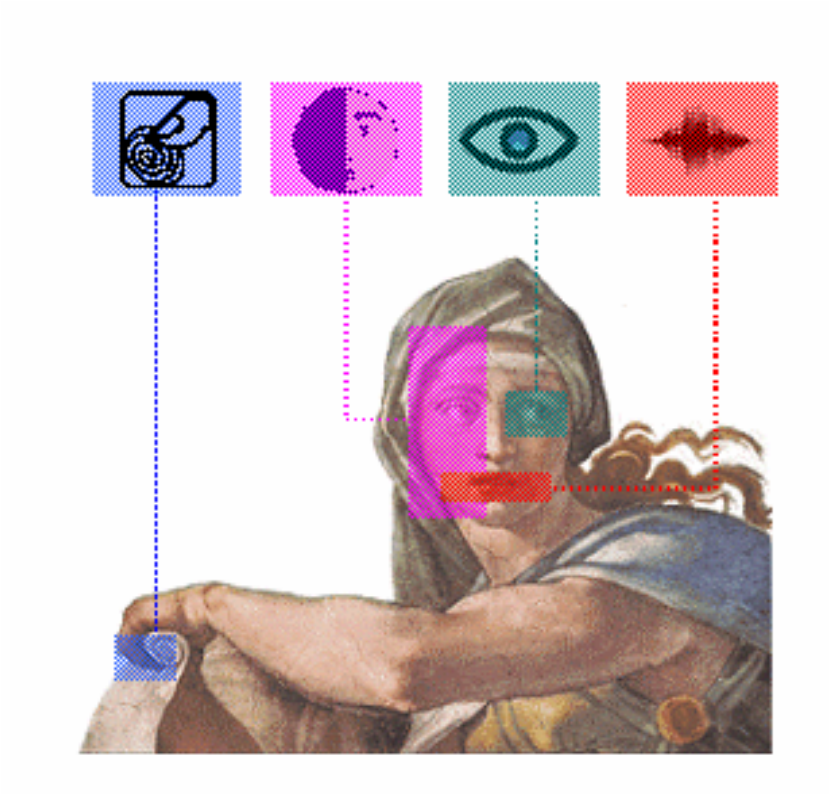
European Co-Operation in the Field
of Scientific and Technical Research

Telecommunications, Information
Science & Technology

Third COST 275 Workshop

# Biometrics on the Internet

University of Hertfordshire, UK
27-28 October 2005



# PROCEEDINGS

# COST

**European cooperation in the field of scientific and technical research**

COST Action 275

## PROCEEDINGS OF THE THIRD COST 275 WORKSHOP

# Biometrics on the Internet

***Edited by***

**Aladdin Ariyaeeinia, Mauro Falcone and Andrea Paoloni**

## LEGAL NOTICE

**By the COST Office**

Third COST 275 Workshop

# Biometrics on the Internet

University of Hertfordshire
Hatfield, UK
27-28 October 2005

**ORGANISERS:**

## University of Hertfordshire

## COST Action 275

## Third COST 275 Workshop

# Biometrics on the Internet

### Sponsors/Supporters:

**COST – TIST**
European Co-Operation in the Field of Scientific and Technical Research, Telecommunications, Information Science and Technology


**University of Hertfordshire**


**IEE/PN-VIE**
Institution of Electrical Engineers/ Professional Network – Visual Information Engineering


**IEEE UKRI CAS**
Institute of Electrical and Electronic Engineers


**EBF**
European Biometrics Forum

**FUB**
Fondazione Ugo Bordoni

**ISCA**
International Speech Communication Association


**SPLC**
Speaker and Language Characterisation – A special interest group of ISCA

## Third COST 275 Workshop

# Biometrics on the Internet

## PROGRAMME COMMITTEE

| | |
|---|---|
| **Aladdin Ariyaeeinia** | University of Hertfordshire, UK |
| **Jean F. Bonastre** | University of Avignon, France |
| **Herve Bourlard** | IDIAP, Switzerland |
| **Andrzej Drygajlo** | EPFL, Switzerland |
| **Mauro Falcone** | FUB, Italy |
| **Carmen Garcia-Mateo** | University of Vigo, Spain |
| **Josef Kittler** | University of Surrey, UK |
| **Javier Ortega Garcia** | Universidad Autonoma de Madrid, Spain |
| **Andrea Paoloni** | FUB, Italy |
| **Nikola Pavesić** | University of Ljubljana, Slovenia |
| **Ioannis Pitas** | University of Thessaloniki, Greece |

## SCIENTIFIC COMMITTEE

| | |
|---|---|
| **Laurent Besacier** | CLIPS - Universite Joseph Fourier, France |
| **Jean F. Bonastre** | University of Avignon, France |
| **Anton Cizmar** | Technical University of Kosice, Slovakia |
| **Farzin Deravi** | University of Kent, UK |
| **Andrzej Drygajlo** | EPFL, Switzerland |
| **Michael Fairhurst** | University of Kent, UK |
| **Carmen Garcia-Mateo** | University of Vigo, Spain |
| **Jozef Juhar** | Technical University of Kosice, Slovakia |
| **Josef Kittler** | University of Surrey, UK |
| **Constantine Kotropoulos** | Aristotle University of Thessaloniki, Greece |
| **Sebastien Marcel** | IDIAP, Switzerland |
| **Håkan Melin** | KTH, Sweden |
| **Javier Ortega-Garcia** | Universidad Autonoma de Madrid, Spain |
| **Nikola Pavesić** | University of Ljubljana, Slovenia |
| **Slobodan Ribarić** | University of Zagreb, Croatia |
| **Piotr Staroniewicz** | Wroclaw University of Technology, Poland |
| **Johann Siau** | University of Hertfordshire, UK |

## ORGANISING COMMITTEE

| | |
|---|---|
| **Aladdin Ariyaeeinia** | University of Hertfordshire, UK |
| **Mauro Falcone** | FUB, Italy |
| **Anthony Herbland** | University of Hertfordshire, UK |
| **Johann Siau** | University of Hertfordshire, UK |

# FOREWORD

**The past few years have witnessed an unprecedented level of acceleration in research into biometrics and its deployment in a variety of scenarios and applications. The impetus for this has been based on a combination of two factors. Firstly, there has been an ever-growing demand for reliable methods for automatic authentication of personal identities. Secondly, the favourable characteristics and potential benefits of biometrics have made it a promising candidate for this purpose.**

**Since 2001, COST 275* has made significant contributions to this area by operating as a valuable platform for the European research collaboration in biometrics and its delivery over the Internet. The Action has involved over thirty participating establishments from fourteen countries: Belgium, Greece, Croatia, France, Ireland, Italy, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden, Switzerland and United Kingdom.**

**The activities within COST 275, over the last four years, have covered a wide range of areas including unimodal biometrics (voice and image-based biometrics); fusion methods for multimodal biometrics; architectures for online biometrics; and assessment means. During the course of its activities, the Action has had a key role in accelerating advances in the above areas and in facilitating future work and progress. Moreover, the Action has exhibited a consistent commitment to effective dissemination of research results with the aim of broadening awareness of the field. Various approaches have been adopted for this purpose including publications of technical reports and research papers as well as major contributions to the relevant special issues of academic journals. One of the most successful means of disseminating the Action outputs has undoubtedly been the organisation of workshops and similar events. The Third COST 275 workshop is in fact the last in this series of events, with the additional aim of marking the conclusion of the Action activities.**

**The scope of the Workshop covers a wide range of topics in the field. These include speaker verification and identification, face recognition and other classes of image-based biometrics, fusion methods and multimodality, robustness against degradation in data, operating conditions and implementation issues. The Technical Programme includes a number of keynote presentations by renowned researchers in the field and four regular sessions covering the contributed papers.**

**We would like to express our appreciation to all the sponsors and supporters of the Workshop and, in particular, COST-TIST, COST Office, IEE, IEEE (UK Section), EBF (European Biometrics Forum), ISCA (International Speech Communication Association), University of Hertfordshire, and FUB (Fondazione Ugo Bordoni). We are also grateful to all the authors and co-authors of the keynote and regular papers for their valuable contributions. Finally, we would to express our gratitude to the members of the Scientific Committee for reviewing the paper proposals in a relatively short period of time, and for their efforts in ensuring high standards for the technical programme.**

**Aladdin Ariyaeeinia**
**Mauro Falcone**
**Andrea Paoloni**

---

*** Further information on COST Action 275 can be found at: http://www.fub.it/cost275**

## Third COST 275 Workshop

# Biometrics on the Internet

## KEYNOTE SPEECHES

1. **The Role of the EBF in the EU Research Arena**
   **Max Snijder, CEO, European Biometrics Forum (EBF)**

2. **Towards Novel Facial Biometrics**
   **Jean-Luc Dugelay, EURECOM, France**

3. **Challenges of Large Scale Speaker Recognition**
   **Homayoon Beigi, Recognition Technologies Inc, USA**

4. **Multiple Expert Fusion and Information Fusion in Biometrics**
   **Fabio Roli, University of Cagliari, Italy**

5. **Finding Differences Between Faces: A Behavioural Approach**
   **Enrico Grosso, University of Sassari, Italy**

# TABLE OF CONTENTS

## IMAGE-BASED BIOMETRICS

## SPEAKER RECOGNITION

## MULTIMODALITY AND EMERGING TECHNOLOGIES

## SYSTEMS AND APPLICATIONS

# AUTHOR INDEX

# IMAGE-BASED BIOMETRICS

# A REVIEW OF SCHEMES FOR FINGERPRINT IMAGE QUALITY COMPUTATION

*Fernando Alonso-Fernandez, Julian Fierrez-Aguilar, Javier Ortega-Garcia*

Biometrics Research Lab.- ATVS, Escuela Politecnica Superior - Universidad Autonoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - Campus de Cantoblanco - 28049 Madrid, Spain
email: {fernando.alonso, julian.fierrez, javier.ortega}@uam.es

**ABSTRACT**

Fingerprint image quality affects heavily the performance of fingerprint recognition systems. This paper reviews existing approaches for fingerprint image quality computation. We also implement, test and compare a selection of them using the MCYT database including 9000 fingerprint images. Experimental results show that most of the algorithms behave similarly.

## 1. INTRODUCTION

Due to its permanence and uniqueness, fingerprints are widely used in many personal identification systems. Fingerprints are being increasingly used not only in forensic environments, but also in a large number of civilian applications such as access control or on-line identification [1].

The performance of a fingerprint recognition system is affected heavily by fingerprint image quality. Several factors determine the quality of a fingerprint image: skin conditions (e.g. dryness, wetness, dirtiness, temporary or permanent cuts and bruises), sensor conditions (e.g. dirtiness, noise, size), user cooperation, etc. Some of these factors cannot be avoided and some of them vary along time. Poor quality images result in spurious and missed features, thus degrading the performance of the overall system. Therefore, it is very important for a fingerprint recognition system to estimate the quality and validity of the captured fingerprint images. We can either reject the degraded images or adjust some of the steps of the recognition system based on the estimated quality.

Fingerprint quality is usually defined as a measure of the clarity of ridges and valleys and the "extractability" of the features used for identification such as minutiae, core and delta points, etc [2]. In good quality images, ridges and valleys flow smoothly in a locally constant direction [3].

In this work, we review the algorithms proposed for computing fingerprint image quality. We also implement, test and compare a selection of them using the MCYT database [4, 5].

The rest of the paper is organized as follows. We review existing algorithms for fingerprint image quality computation in Sect. 2. An experimental comparison between selected techniques is reported in Sect. 3. Conclusions are finally drawn in Sect. 4.

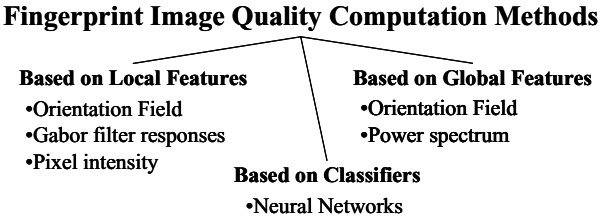



**Fig. 1**. A taxonomy of fingerprint image quality computation algorithms.

## 2. FINGERPRINT IMAGE QUALITY COMPUTATION

A taxonomy of existing approaches for fingerprint image quality computation is shown in Fig. 1. We can divide the existing approaches into $i$) those that use local features of the image; $ii$) those that use global features of the image; and $iii$) those that address the problem of quality assessment as a classification problem.

### 2.1. Based on local features

Methods that rely on local features [2, 3, 6-8] usually divide the image into non-overlapped square blocks and extract features from each block. Blocks are then classified into groups of different quality. A *local measure of quality* is finally generated. This local measure can be the percentage of blocks classified with "good" or "bad" quality, or an elaborated combination. Some methods assign a relative weight to each block based on its distance from the centroid of the fingerprint image, since blocks near the centroid are supposed to provide more reliable information [2, 8].

#### 2.1.1. Based on the orientation field

This group of methods use the local angle information provided by the orientation field to compute several local features in each block. Hong et al. [3] modeled ridges and valleys as a sinusoidal-shaped wave along the direction normal to the local ridge orientation and extracted the amplitude, frequency and variance of the sinusoid. Based on these parameters, they classify the blocks as *recoverable* and *unrecoverable*. If the percentage of unrecoverable blocks exceeds a predefined threshold, the image is rejected. The method presented by Lim et al. [6] computes the following features in each block: orientation certainty level, ridge frequency, ridge thickness and ridge-to-valley thickness ratio. Blocks are then labeled as "good", "undetermined", "bad" or "blank" by thresholding the four local features. A local quality score $S_L$ is computed based on the total number of "good", "undetermined" and "bad" quality image blocks. Recently, Chen et al. [2] proposed a local quality index which measures the spatial coherence using the intensity gradient. The orientation coherence in each block is computed. A local quality score $Q_S$ is finally computed by averaging the coherence of each block, weighted by its distance to the centroid of the foreground.

#### 2.1.2. Based on Gabor filters

Shen et al. [7] proposed a method based on Gabor features. Each block is filtered using a Gabor filter with $m$ different orientations. If a block has good quality (i.e. strong ridge orientation), one or several filter responses are larger than the others. In poor quality blocks or background blocks, the $m$ filter responses are similar. The standard deviation of the $m$ filter responses is then used to determine the quality of each block ("good" and "poor"). A quality index $QI$ of the whole image is finally computed as the percentage of foreground blocks marked as "good". If $QI$ is lower than a predefined threshold, the image is rejected. Poor quality images are additionally categorized as "smudged" or "dry".

#### 2.1.3. Based on pixel intensity

The method described in [8] classifies blocks into "directional" and "non-directional" as follows. The sum of intensity differences $D_d\,(i,j)$ between a pixel $(i,j)$ and $l$ pixels selected along a line segment of orientation $d$ centered around $(i,j)$ is computed for $n$ different orientations. For each different orientation $d$, the histogram of $D_d\,(i,j)$ values is obtained for all pixels within a given foreground block. If only one of the $n$ histograms has a maximum value greater than a prominent threshold, the block is marked as "directional". Otherwise, the block is marked as "non-directional".

An overall quality score $Q$ is finally computed. A relative weight $w_i$ is assigned to each foreground block based on its distance to the centroid of the foreground. $Q$ is defined as $Q = \sum_D w_i \,/\, \sum_F w_i$ where $D$ is the set of directional blocks and $F$ is the set of foreground blocks. If $Q$ is lower than a threshold, the image is considered to be of poor quality. Measures of the smudginess and dryness of poor quality images are also defined.

### 2.2. Based on global features

Methods that rely on global features [2, 6] analyze the overall image and compute a *global measure of quality* based on the features extracted.

#### 2.2.1. Based on the orientation field

Lim et al. [6] presented two features to analyze the global structure of a fingerprint image. Both of them use the local angle information provided by the orientation field, which is estimated in non-overlapping blocks. The first feature checks the continuity of the orientation field. Abrupt orientation changes between blocks are accumulated and mapped into a global orientation score $S_{GO}$. The second feature checks the uniformity of the frequency field [9]. This is done by computing the standard deviation of the ridge-to-valley thickness ratio and mapping it into a global score $S_{GR}$. Although ridge-to-valley thickness is not constant in fingerprint images in general, the separation of ridges and valleys in good quality images is more uniform than in low quality ones. Thus, large deviation indicates low image quality.

#### 2.2.2. Based on Power Spectrum

Global structure is analyzed in [2] by computing the 2D Discrete Fourier Transform (DFT). For a fingerprint image, the ridge frequency value lies within a certain range. A region of interest (ROI) of the spectrum is defined as an annular region with radius ranging between the minimum and maximum typical ridge frequency values. As fingerprint image quality increases, the energy will be more concentrated in ring patterns within the ROI. The global quality index $Q_F$ defined in [2] is a measure of the energy concentration in ring-shaped regions of the ROI. For this purpose, a set of bandpass filters is constructed and the amount of energy in ring-shaped bands is computed. Good quality images will have the energy concentrated in few bands.

### 2.3. Based on classifiers

The method that uses classifiers [10] defines the quality measure as a degree of separation between the match and non-match distributions of a given fingerprint. This can be seen as a prediction of the matcher performance.

*2.3.1. Based on neural networks*

Tabassi et al. [10] presented a novel strategy for estimating fingerprint image quality. They first extract the fingerprint features used for identification and then compute the quality of each extracted feature to estimate the quality of the fingerprint image, which is defined as the degree of separation between the match and non-match distributions of a given fingerprint.

Let $s_m(x_i)$ be the similarity score of a genuine comparison (*match*) corresponding to the subject $i$, and $s_n(x_{ji})$, $i \neq j$ be the similarity score of an impostor comparison (*non-match*) between subject $i$ and impostor $j$. Quality $Q_N$ of a biometric sample $x_i$ is then defined as the prediction of

$$o(x_i) = \frac{s_m(x_i) - E[s_n(x_{ji})]}{\sigma(s_n(x_{ji}))} \quad (1)$$

where $E[.]$ is mathematical expectation and $\sigma(.)$ is standard deviation. Eq. 1 is a measure of separation between the *match* and the *non-match* distributions, which is supposed to be higher as image quality increases.

The prediction of $o(x_i)$ is done in two steps: $i)$ $v_i = L(x_i)$; and $ii)$ $Q_N = \tilde{o}(x_i) = I(v_i)$; where $L(.)$ computes a feature vector $v_i$ of $x_i$ and $I(.)$ maps the feature vector $v_i$ to a prediction $\tilde{o}(x_i)$ of $o(x_i)$ by using a neural network.

Feature vector $v_i$ contains the following parameters: $a)$ number of foreground blocks; $b)$ number of minutiae found in the fingerprint; $c)$ number of minutiae that have quality value higher than 0.5, 0.6, 0.75, 0.8 and 0.9, respectively; and $d)$ percentage of foreground blocks with quality equal to 1, 2, 3 and 4, respectively. All those values are provided by the MINDTCT package of NIST Fingerprint Image Software (NFIS) [11]. This method uses both local and global features to estimate the quality of a fingerprint.

## 3. EXPERIMENTS

In this work, we have implemented an tested some of the algorithms presented above using the existing fingerprint image database MCYT [4, 5]. In particular, 9000 fingerprint images from all the fingers of 75 subjects are considered (QMCYT subcorpus from now on). Fingerprints are acquired with an optical sensor, model UareU from Digital Persona, with a resolution of 500 dpi and a size of 400 pixels height and 256 pixels width. A subjective quality assessment $Q_M$ of this database was accomplished by a human expert. Each different fingerprint image has been assigned a subjective quality measure from $0$ (lowest quality) to $9$ (highest quality) based on factors like: captured area of the fingerprint, pressure, humidity, amount of dirt, and so on.

The algorithms tested in this work are as follows: $i)$ the combined quality measure $Q_C$ computed in [6] by linearly combining the scores $S_L$, $S_{GO}$ and $S_{GR}$ presented in Sects. 2.1.1 and 2.2.1; $ii)$ the algorithms presented in [2] based on local $Q_S$ (Sect. 2.1.1) and global features $Q_F$ (Sect. 2.2.2); and $iii)$ the method $Q_N$ based on neural networks proposed in [10] and described in Sect. 2.3.1. The quality measures $Q_C$, $Q_S$ and $Q_F$ lie in the range $[0, 1]$ whereas $Q_N \in \{1, 2, 3, 4, 5\}$. The selected methods are also compared with the subjective quality assessment $Q_M$ accomplished in QMCYT.

**Fig. 2**. Sample images extracted from the five quality subsets created using the manual quality measure $Q_M$. Images are arranged by increasing quality (on the left: lowest quality, subset 1; on the right: highest quality, subset 5).

The above-mentioned quality measures have been computed for all QMCYT. In order to compare the selected methods, we have arranged the fingerprint images by increasing quality measure $Q_k$, $k \in \{M, N, C, S, F\}$. Then, 5 subsets $S_k^i$, $i \in \{1, 2, 3, 4, 5\}$, of equal size (1800 images per subset) are created. The first subset contains the 1800 images with the lowest quality measures, the second subset contains the next 1800 images with the lowest quality measures, and so on. Sample images extracted from the five quality subsets created using the manual quality measure $Q_M$ are shown in Fig. 2. The mean quality measure of each subset $S_k^i$ is then computed as $\tilde{Q}_k^i = \frac{1}{1800} \sum_{j \in S_k^i} Q_k^i(j) n(j)$ where $n(j)$ is the total number of images with quality measure $Q_k^i(j)$. Lastly, mean quality measures $\tilde{Q}_k^i$ are normalized to the $[0, 1]$ range as follows: $\hat{Q}_k^i = \left(\tilde{Q}_k^i - \tilde{Q}_k^1\right) / \left(\tilde{Q}_k^5 - \tilde{Q}_k^1\right)$ where $\hat{Q}_k^i$ is the normalized mean quality measure of $\tilde{Q}_k^i$.

In Fig. 3, we can see the normalized mean quality measures $\hat{Q}_k^i$ of each subset $S_k^i$, $i \in \{1, 2, 3, 4, 5\}$, for all the $k$ algorithms tested, $k \in \{M, N, C, S, F\}$. Maximum value, minimum value and standard deviation value of normalized individual quality measures of each subset are also depicted. It can be observed that that most of the algorithms result in similar behavior, assigning well-separated quality measures to different quality groups. Only the algorithm based on classifiers, $Q_N$, results in very different behavior, assigning the highest quality value to more than half of the database. It may be due to the low number of quality labels used by this algorithm [10].

Regarding to the algorithms that behave similarly, it can be observed that standard deviation is similar for quality groups 2 to 4. Only the method based on the subjective quality assessment $Q_M$ results in slightly higher deviation. This

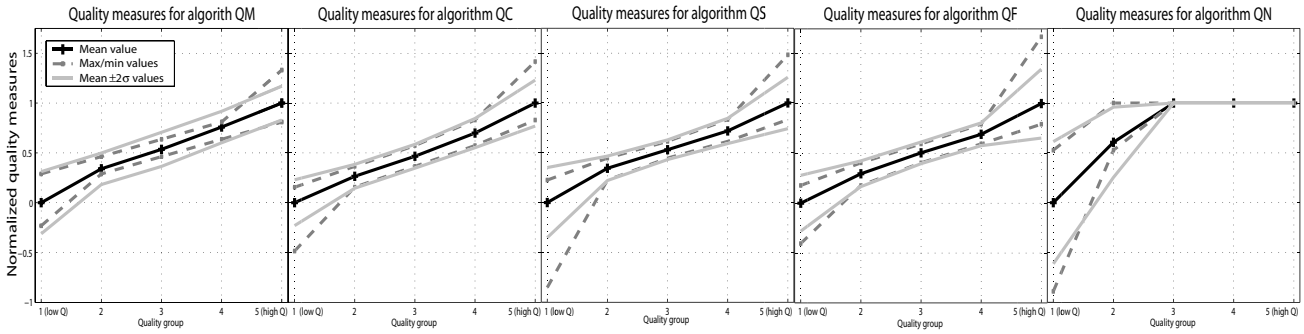



**Fig. 3**. Normalized mean quality measure $\hat{Q}_k^i$ of quality group $i \in \{1, 2, 3, 4, 5\}$, for all the $k$ algorithms tested (M=Manual, C=Combined local+global features [6], S=local spatial features [2], F=global frequency [2], N=classifier based on neural networks [10]). Maximum value, minimum value and standard deviation value of normalized quality measures of each quality group are also depicted.

is maybe due to the finite number of quality labels used. The other algorithms assign continuous quality measures within a certain range.

In addition, in most of the quality groups, normalized quality measures lie within a range of 2 times the standard deviation. Only quality groups 1 and 5 sometimes behave different, maybe to the presence of outliers (i.e., images with very low quality measure in group 1 and with very high quality measure in group 5, respectively).

### 4. CONCLUSIONS AND FUTURE RESEARCH

This paper reviews most of the existing algorithms proposed to compute the quality of a fingerprint image. They can be divided into $i)$ those that use local features of the image; $ii)$ those that use global features of the image; and $iii)$ those that address the problem of quality assessment as a classification problem. We have implemented and tested a selection of them. They are compared with the subjective quality assessment accomplished in the existing QMCYT subcorpus. Experimental results show that most of the algorithms behave similarly, assigning well-separated quality measures to different quality groups. Only the algorithm based on classifiers [10] results in very different behavior. It may be due to the low number of quality labels used by this algorithm. Future work includes integrating the implemented quality estimation algorithms into a quality-based multimodal authentication system [12].

## Acknowledgments

This work has been supported by BioSecure European NoE and the TIC2003-08382-C05-01 project of the Spanish Ministry of Science and Technology. F. A.-F. and J. F.-A. thank Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting their PhD studies.

### 5. REFERENCES

[1] A. K. Jain, A. Ross and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Tech.*, 14(1):4–20, January 2004.

[2] Y. Chen et al. Fingerprint quality indices for predicting authentication performance. *Proc. AVBPA - to appear*, 2005.

[3] L. Hong et al. Fingerprint image enhancement: Algorithm and performance evaluation. *IEEE Trans. on PAMI*, 20(8):777–789, August 1998.

[4] J. Ortega-Garcia, J. Fierrez-Aguilar et al. MCYT baseline corpus: a bimodal biometric database. *IEE Proc. VISP*, 150(6):395–401, December 2003.

[5] D. Simon-Zorita, J. Ortega-Garcia et al. Image quality and position variability assessment in minutiae-based fingerprint verification. *IEE Proc. VISP*, 150(6):402-408, Dec. 2003.

[6] E. Lim et al. Fingerprint quality and validity analysis. *IEEE Proc. ICIP*, 1:469–472, September 2002.

[7] L. Shen et al. Quality measures of fingerprint images. *Proc. AVBPA*: 266–271, 2001.

[8] N. Ratha and R. Bolle (Eds.). *Automatic Fingerprint Recognition Systems*. Springer-Verlag, N.York, 2004.

[9] D. Maltoni, D. Maio, A. Jain and S.Prabhakar. *Handbook of Fingerprint Recognition*. Springer, N.York, 2003.

[10] E. Tabassi, C. Wilson and C. Watson. Fingerprint image quality. *NIST research report NISTIR7151*, 2004.

[11] C.I. Watson et al. *User's Guide to Fingerprint Image Software 2 - NFIS2*. NIST, 2004.

[12] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez and J. Bigun. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition*, 38(5):777–779, May 2005.

# ROBUST METHOD OF REFERENCE POINT LOCALIZATION IN FINGERPRINTS

*Krzysztof Kryszczuk and Andrzej Drygajlo*

Speech Processing and Biometrics Group

Signal Processing Institute, Swiss Federal Institute of Technology Lausanne, Switzerland

{krzysztof.kryszczuk, andrzej.drygajlo}@epfl.ch

## ABSTRACT

Successful and robust location of reference point or points in a fingerprint is a key element of many fingerprint recognition systems. Traditionally, singular points in a fingerprint are used as such points. However, incomplete capture, deteriorated quality and inherent fingerprint characteristics can make a reliable singularity detection impossible. In this paper we propose the use of multiple reference points, and we present a new method of localizing them. The advantage of the proposed method is that it allows for successful detection and prioritization of multiple reference points in a fingerprint, and that it uses the same features that can be subsequently used for recognition.

## 1. INTRODUCTION

The use of biometrics on low-power mobile, hand-held devices and over wireless networks forces new range of constraints and requirements on both data acquisition and its processing. Successful deployment of a fingerprint recognition system in a mobile scenario requires that the applied algorithms be able to cope with many adverse fingerprint capture conditions. In particular, lowered fingerprint quality and incomplete capture deserve special attention

The vast majority of contemporary automated fingerprint authentication systems are minutiae (level 2 features) based systems [5,7,8]. Minutiae-based systems normally perform well with high-quality fingerprint images and a sufficient fingerprint surface area. These conditions, however, may not always be attainable. In many cases, only a small portion of the test fingerprint can be compared with the reference fingerprint. Comparing fingerprints acquired through small-area sensors is difficult due to the possibility of having too little overlap between different acquisitions of the same finger [5]. If a minutiae-based matching algorithm is used, in the case of small overlap between the fingerprints, the number of minutiae correspondences might significantly decrease and the matching algorithm would not be able to make a decision with high certainty. This effect is even more marked on intrinsically poor quality fingers, where only a subset of the minutiae can be extracted and used with sufficient reliability.

Recently we have proposed a fingerprint matching scenario where, in addition to the minutiae, sweat pores are used [3]. The proposed algorithm offers increased robustness to incomplete capture, but it does not address the problems of low-quality fingerprints. Also, reliable estimation of sweat pore loci depends heavily on the capture resolution. It is necessary that the sensor is able to deliver fingerprints at the resolution of at least 1000 dpi [3], while the resolution of the devices available on the market does not usually exceed 550 dpi.

Minutiae-based fingerprint matching can be considered to be a link between traditional, manual matching routines, and today's need for fully automatic systems. Jain et al [2] proposed a FingerCode-based scheme of fingerprint matching, which was shown to outperform a minutiae-based matcher. The advantage of the FingerCode-based fingerprint matching is its robustness to low quality fingerprints. It also does not depend on the number of minutiae present in the fingerprint area used for matching. The FingerCode implicitly captures both characteristic ridge formations (minutiae) as well as the ridge flow and pattern itself, it therefore does not a priori require any number of minutiae to be present.

The critical part of the FingerCode-based fingerprint recognition scheme is the localization of the reference point, from which the extraction of the FingerCode begins. The subject literature reports many ways of localizing the reference point in the fingerprint. Typically the singular point or points in a fingerprint is used as such a reference [1,5].

There are rarely more than 2 singular points in an impression of a fingerprint, hence a strict use of singular points as reference points for feature extraction is a potential source of major recognition errors. There is nothing inherently special about singular points that would predestine them to be reference points. Therefore we argue in this paper that a robust and consistent detection of multiple reference points in a fingerprint can lead to improved recognition accuracy.

Most of the algorithms found in the literature process the fingerprint orientation map [5] to find singular points in fingerprints. Estimation of the orientation map of a fingerprint

and its smoothing is a nontrivial task if the impression is of even slightly deteriorated quality, and imposes an additional computational burden.

In this paper, we present a new algorithm for detecting reference points in fingerprints. The algorithm does not require that the orientation map of the fingerprint be created. Instead, it uses the image responses to Gabor filters, which can be directly used as features for recognition. The method performs well even when the fingerprint quality is strongly deteriorated.

In our work we used fingerprints from the publicly available databases from FVC2000 [4] and FVC 2002.

## 2. THE CHOICE OF A REFERENCE POINT

Traditionally, singular points have been used as reference points since they are also used for indexing and searching through large fingerprint databases [1,6,7]. To human observers, singular points are the obvious choice as a reference point since they are easy to recognize just by observing the ridge pattern. From the viewpoint of automatic fingerprint recognition though there is no reason to choose the singular points over any other point of the fingerprint. However, the neighborhood of the singular point is likely to be captured in a fingerprint scanner, and it contains many characteristic features, making both reference point localization and fingerprint recognition possible. Such a feature is the ridge orientation. Similarly as it is the case with the orientation map-based methods, we make use of the information contained in the local ridge orientation to locate the points of reference. However, we do not employ the constraint that the reference point should be a singular point of the fingerprint. In this way, we can afford to detect multiple reference points.

The proposed method consists of following steps:

1. Fingerprint cropping/boundary detection
2. Building a ridge orientation image by filtering the image with a bank of Gabor filters.
3. Localizing the reference points.

## 3. FINGERPRINT CROPPING

Fingerprint cropping is a process of selecting the image area that are indeed occupied by the clear impression of the ridges, and discarding the background that may contain sensor noise and other artifacts. Discarding irrelevant parts of the image also speeds up the search for reference points.

Before processing, each fingerprint image is mean-subtracted. In order to find the fingerprint boundary the entire image is divided into non-overlapping blocks of 8×8 pixels. For each block, the variance of the local pixel intensity value is computed. Blocks of variance inferior to a selected threshold are discarded as background. The choice of threshold is empirical and depends on the sensor used, but it is not a very sensitive parameter.

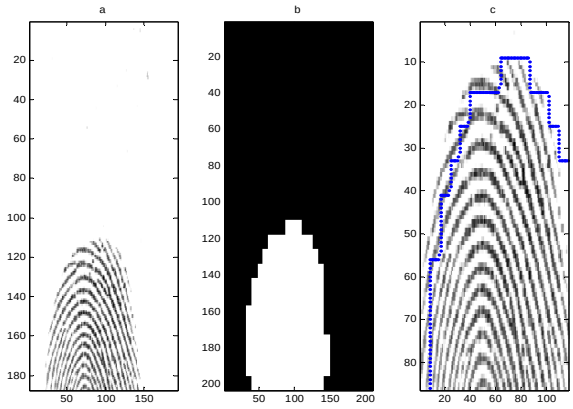



**Figure 1:** Fingerprint cropping: (a) original image from FVC2002, Db1_a, (b) crop mask and (c) cropped image with marked boundary.

## 4. FILTERING

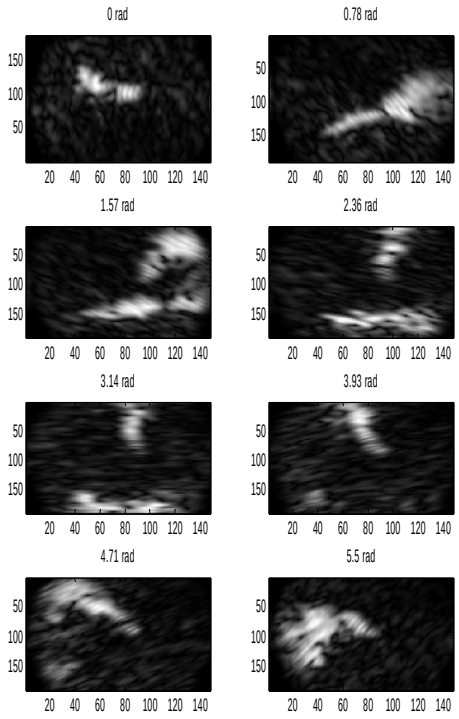



**Figure 2:** Example of a fingerprint filtered with 8 directional filters.

### 4.1 Gabor filtering

In order to obtain information about the local ridge information, the cropped fingerprint image is filtered using a bank of Gabor filters. At least 4 filters are required to capture local ridge information [2]. In our work, however, we used eight filter orientations, because the reference point detection

proved to be more reliable, and the filter responses can be directly used as features in the recognition process [2].

The filter parameters are chosen in order to provide robustness to local fingerprint quality degradation. It is important to select the filter base frequency, σx and σy so that the filters respond to the ridge flow, but not to image noise.

## 4.2 Building an orientation image

The information from the single filter responses are combined to create an orientation image of a fingerprint. Since we are interested in the dominant local orientation and not in the magnitude of single filter responses, for each pixel of the image we assign the orientation of the filter whose response is strongest.

## 4.3 Cleaning the orientation image

Despite the fact that due to a proper parameter choice the noise impact on the local ridge orientation estimation is suppressed, some local ridge flow irregularities do show in the orientation image. Since those irregularities do not contribute to the reference point localization process, they need to be cleaned.

The cleaning of the orientation image is done as follows:

- Each of the patches of uniform orientation is labeled
- For each labeled patch, its area is computed. If the patch area is smaller than a selected threshold, the orientation assigned to all pixels of the patch is changed to a new orientation.
- The new orientation of the patch is chosen according to the dominant orientation of the immediate neighborhood of the patch.

Since eliminating certain small matches leads to creation of new orientation image, which may still contain patches whose area is below the selected threshold, the above procedure is

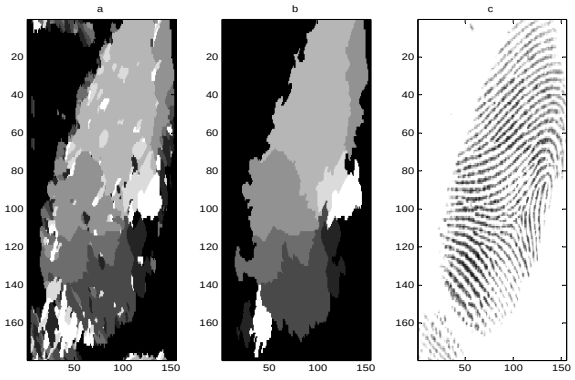



**Figure 3:** Fingerprint (c) orientation image before (a) and after cleaning (b)

repeated until areas of all patches in the orientation image are above the threshold.

# 5. LOCALIZING REFERENCE POINTS

It is apparent from Figure 3 that the boundaries dividing the patches of different ridge orientations in the cleaned orientation image converge towards the neighborhood of the singularities. Due to the characteristic ridge flow and to the dynamic deformation from one imprint of a finger to another it is often difficult to accurately pinpoint the actual singularity. We propose to correct in original the criteria of nominating candidates for reference points. Instead of insisting on finding a northernmost point on the ridges where the ridge curvature is maximal [5], we attempt to find the points of convergence of the boundaries between different orientations in the orientation image. This approach can be classified as orientation field segmentation [5]. However computing only the segmentation lines looses the information on which orientations actually converged into a given point. In noisy fingerprints, certain areas can have segmentation lines accidentally convergent to points which are not singular. To avoid this problem, instead of counting the number of segmentation boundary lines converging to a point, we count the number of different orientations of patches that are adjacent to the point.

In general, the more various orientations surround given point, the more certain we are that given point is a robustly localized reference point. The number of orientations present in an immediate neighborhood of a point can be thus treated as a measure of certainty that given reference point was correctly detected. The proposed method localizes and assigns the certainty measures to reference points iteratively. First, the orientation image is downsampled since looking for candidates at each pixel is not necessary. Then, for every pixel in the orientation image we calculate the number of different orientations present in an 11×11 neighborhood. The choice of the neighborhood is empirical; an increased neighborhood increases the chances of localizing the reference point at the cost of precision. If $n$ is the total number of Gabor filters, then only those points in whose neighborhood at least $n$+1 orientations are present are considered reference points. The number of orientations present in the neighborhood of the localized reference point constitutes a measure of confidence in the correct detection of the reference point.

# 6. EXPERIMENTAL RESULTS

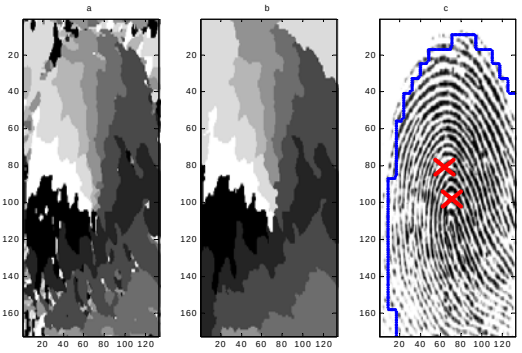



**Figure 4:** Localized reference points in a fingerprint (high quality impression)

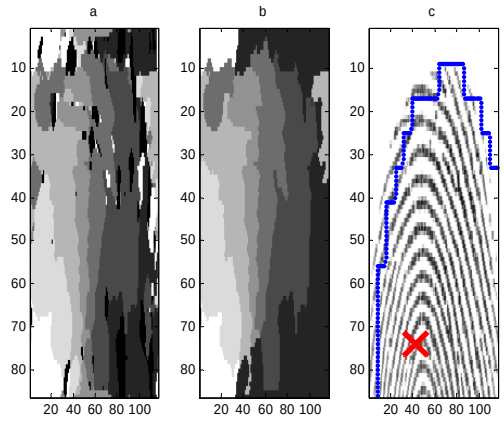



**Figure 5:** Localized reference points in an example fingerprint (fragmentary impression)

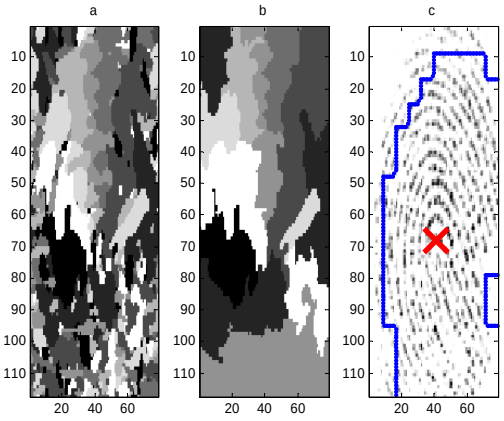



**Figure 6:** Localized reference points in an example fingerprint (low quality impression)

Figures 4,5 and 6 show an example of successful reference point localization in high quality, fragmentary, and low quality impressions of the same fingerprint. While two reference points are localized in the high quality sample (presumably a fingerprint that in a real system could be used as a training sample), only one reference point could be located in fragmentary and low quality impression. In a system that uses a single reference point such a situation could lead to erratic recognition.

Figures 4, 5 and 6: a. orientation image, b. cleaned orientation image, c. cropped fingerprint with marked reference points.

## 7. CONCLUSIONS

We have proposed a new method for reference point extraction in fingerprint. The presented method uses a liberal definition of a reference point and does not require it to be positioned exactly at the singular point of the fingerprint. Rather than one singular point, the output of the proposed scheme is a set of reference points. The use of multiple reference points maximizes the chances that in the presence of fingerprint quality deterioration one of the detected reference points will correspond to one of the reference points in the training (gallery) fingerprints. In this sense, the proposed method could be viewed as a compromise between a computationally expensive exhaustive search for a match of the feature template in test fingerprint, and one-to-one template matching, which is sensitive to errors in reference point localization.

As presented in this paper, the method does not differentiate between different types of singular points (core, delta etc.). From the recognition point of view making such a distinction is unnecessary. Also, multiple reference points cannot by definition all be singular. They, however, usually are positioned in the vicinity of a singularity, and the proximity of a singularity can provide an appropriate label for a reference point, if necessary. A computationally inexpensive method of labeling the reference points is to filter its neighborhood in a manner similar to that presented in [6]. The proposed method is computationally inexpensive since it employs the same features that can be reused in the verification process.

## 8. REFERENCES

1. Bazen, A.M.: Fingerprint Identification. Feature Extraction, Matching and Database Search. Doctoral Thesis, Twente University, Enschede, 2002.

2. Jain, A., Prabhakar, Hong L.: FingerCode: A Fiterbank for Fingerprint Representation and Matching, 1999 IEEE CS Conference on Computer Vision and Pattern Recognition (CVPR'99) - Volume 2 p. 2187.

3. Kryszczuk, K.M., Morier, P., Drygajlo, A.: Extraction of Level 2 and Level 3 features for fragmentary fingerprints, submitted to: 2nd COST275 Workshop (Biometrics on the Internet), Vigo, 2004.

4. Maio, D., Maltoni, D., Capelli, R., Waylam, J.L., Jain, A.K.. FVC2000: Fingerprint Verification Competition. In: IEEE Transactions on PAMI, Vol. 24, No. 3, 2002.

5. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition, Springer, New York, 2003.

6. Nilsson K, Bigun J.: Complex Filters Applied to Fingerprint Images Detecting Prominent Symmetry Points Used for Alignment, proc. Workshop on Biometric Authentication (ECCV2002), LNCS 2359, pp. 39-47, New York 2002

7. Pankanti, S., Prabhakar, S., Jain, A.K.: On the Individuality of Fingerprints. In: Proc. of the IEEE Comp. Soc. Conference on Computer Vision and Pattern Recognition, Hawaii, 2001.

8. Roddy, A.R., Stosz J.D.: Fingerprint Features – Statistical Analysis and System Performance Estimates. In: Proceedings of the IEEE, vol. 85, no. 9, pp. 1390-1421, 1997.

# BIOMETRIC IDENTIFICATION USING PALMPRINT LOCAL FEATURES*

*José García-Hernández and Roberto Paredes*

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia (Spain)
{jgarcia,rparedes}@iti.upv.es

## ABSTRACT

In the networked society there are a great number of systems that need biometric identification, so it has become an important issue in our days. This identification can be based on palmprint features. In this work we present a biometric identification system based on palmprint local features. For this purpose, we have compared the error rate (ER) of 3 different preprocessing methods.

## 1. INTRODUCTION

The biometric automatic identification has become an important issue in our days. *Classical* identification can be divided into two categories: token-based (physical key, cards, etc) and password-based. However, these classical methods have a great number of problems and can be easily broken because tokens and passwords can be forgotten or stolen. Front them, the identification methods based on biometric features are rising nowadays. Biometrics offers an effective approach to identify subjects because it is concerned with the unique, reliable and stable personal physiological features. These features can be: iris [2], fingerprints [5, 17, 15], palmprints [18, 3, 19], hand geometry [6], faces [10, 11, 13], voice [14], etc

Local features with nearest neighbor search and direct voting obtains excellent results for various image classification tasks [13, 14, 7] and has shown good performance in different biometrics systems. Face recognition [13] and speaker recognition [14] have been carried out in the past using local features.

*Work supported by the "Agencia Valenciana de Ciencia y Tecnología (AVCiT)" under grant GRUPOS03/031 and the Spanish Project DPI2004-08279-C02-02

This paper shows a biometric palmprint identification based on palmprint local features. While in section  we present the used approach, in section  we present the used data and the experiments. In section , finally, we present our conclusions.

## 2. PROPOSED APPROACH

### 2.1 Image preprocessing

In our approach, we have considered 3 different preprocessing methods: local equalisation, global equalisation or no equalisation (that is, leave unchanged the image). The used *equalisation* method is called histogram equalisation. In this method the result is obtained using the cumulative density function of the image as a transfer function. The result of this process is that the histogram becomes approximately constant for all gray values. For a given image of size $M \times N$ with $G$ gray levels and cumulative histogram $H(g)$ this transfer function is given in equation 1.

$$T(g) = \frac{G-1}{MN} H(g) \tag{1}$$

While *global equalisation* is the application of the histogram equalisation presented above in the whole image, *local equalisation* is the same method but applied locally. By applied locally means the following. We crop the image, starting in the upper-left corner, with a window of size $v$, such as $v \ll w$. The histogram equalisation function is applied to the cropped image. This process is repeated by moving the crop all over the image and for each one applying the equalisation. In figure 1 examples of both equalisations are shown.

### 2.2 Local Features extraction

Local features extraction technique represents each image by many feature vectors belonging to differ-

**Figure** 1: Equalisation example. From left to right: original image, global equalisation image and local equalisation image.

ent regions of the image. Once the image is preprocessed we select the $n$ pixels with higher information content. For this purpose we have chosen a simple and fast method: the local variance in a small window is measured for each pixel and the $n$ pixels with a greater variance are selected. In our case, these windows have a fixed size of $5 \times 5$.

For each of the selected pixels, a $w^2$-dimensional vector of grey values is first obtained in the preprocessed image by application of a $w \times w$ window around it. The dimension of the resulting vectors is then reduced from $w^2$ to 30 using *Principal Component Analysis* (PCA), thus obtaining a compact local representation of a region of the image. We have reduced the dimension to 30 using PCA due to that value gave us the best performance in most classification tasks in the past. This is illustrated in figure 2.

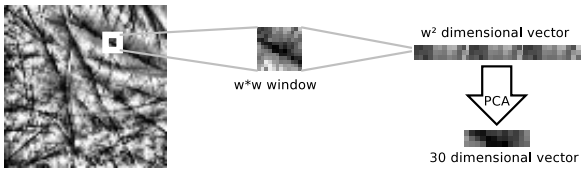



**Figure** 2: Feature extraction process.

## 2.3 Classification through a $k$-NN based Voting Scheme

In a classical classifier, each object is represented by a feature vector, and a discrimination rule is applied to classify a test vector that also represents one object. As discussed above, local representation, however, implies that each image is scanned to compute many feature vectors. Each of them could be classified into a different class, and therefore a decision scheme is required to finally decide a single class for a test image.

Let $Y$ be a test image. Following the conventional probabilistic framework, $Y$ can be optimally classified in a class $\hat{c}$ having the maximum posterior probability among $C$ classes. By applying the feature extraction process described in the previous section to $Y$, a set of $m_Y$ feature vectors, $\{\mathbf{y}_1, \ldots, \mathbf{y}_{m_Y}\}$ is obtained. An approximation to $P(c_j|Y)$ can be obtained using the so called "*sum rule*" and then, the expression of $\hat{c}$ becomes:

$$\hat{c} = \arg\max_{1 \le j \le C} \sum_{i=1}^{m_Y} P(c_j|\mathbf{y}_i) \qquad (2)$$

In our case, posterior probabilities are directly estimated by $k$-Nearest Neighbours. Let $k_{ij}$ the number of neighbours of $\mathbf{y}_i$ belonging to class $c_j$. Using this estimate in (2), our classification rule becomes:

$$\hat{c} = \arg\max_{1 \le j \le C} \sum_{i=1}^{m_Y} k_{ij} \qquad (3)$$

That is, a class $\hat{c}$ with the largest number of *"votes"* accumulated over all the vectors belonging to the test image is selected. This justifies why techniques of this type are often referred to as *"voting schemes"*.

## 2.4 Efficient approximate search of matching feature vectors

An important feature of a practical palmprint recognition method is speed. The preprocessing and feature extraction steps described in section and respectively, have been carefully chosen to be simple and fast, but the retrieval of a large number of high-dimensional vectors for each test image and the search of the $k$-NN among a huge pool of vectors obtained from the reference images seems an intractable issue. A large reference set is essential for the proposed method to be effective. Therefore, a fast search algorithm has to be applied to perform the complete process in a reasonable time. To this end, we have adopted the well known *kd-tree* data structure.

If a guaranteed exact solution is not needed, as can be assumed in our case, the backtracking process associated to the exact *kd-tree* search can be aborted as soon as a certain criterion is met by the current best solution. In [4], the concept of $(1+\epsilon)$-approximate nearest neighbour query is introduced. A point $p$ is a $(1+\epsilon)$-approximate nearest neighbour of $q$ if the distance from $p$ to $q$ is less than $1+\epsilon$ times the distance from $p$ to its nearest neighbour. This concept has been used here to obtain an efficient approximate search that can easily cope with very large sets of reference vectors.

# 3. EXPERIMENTS

In this work we have used the *PolyU Palmprint Database*, created by the Biometric Research Cen-

ter of Hong Kong [1]. This database contains 600 grayscale palmprint images from 100 different palms. There are 6 images for each palmprint and we have selected 3 images for training and 3 for testing. So, both training and testing sets contain 300 palmprint images from 100 different subjects, 3 images from each one. The database is more detailed in [1] and a bigger version is used and described in [19].

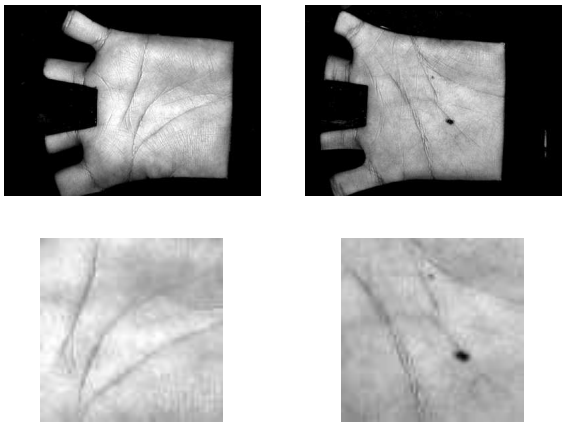


**Figure** 3: Examples of images and its corresponding selection (they are not in the same scale).

The images are correctly placed and neither rotation nor scaling normalization are needed. On the other hand, the here proposed local features approach is invariant to translations. The position of the local features is not stored and then, this position is not used anyway to consider the nearest neighbor of a local feature.

Despite other works [8, 9], where geometrical hand features are also used, we only use the palmprint features. Original database images are $384 \times 284$ sized. After reducing the size to its half, we select an area of $75 \times 75$ pixels. The selected area center is the mass center of the whole hand image, excluding fingers. By this way, we select the palmprint image area with more information. We only work with this selected area and the preprocessing approaches are applied only to this region of interest. Two examples of original images and its corresponding selections are shown in figure 3.

In experiments we have compared the error rate obtained by each of the 3 preprocessing methods showed in section . In all cases we have extracted local features from preprocessed images as it is shown in section  and we have used the voting scheme showed in section . We have varied different parameters values: the local feature window size ($w$), the number of local features extracted for each image ($n$) and in the case of local equalisation also the crop window size ($v$). However, for clarify, in this work we only shown the best results. They are shown in figures 4, 5 and 6. In case of local equalisation, those results are obtained with $v = 9$.

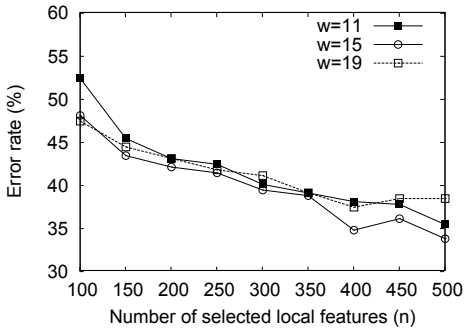



**Figure** 4: Results with no Equalisation preprocessing.

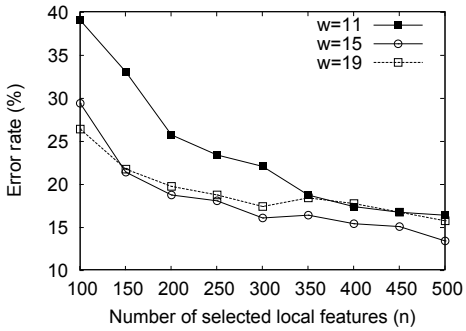



**Figure** 5: Results with Global Equalisation preprocessing.

As can be seen in figures 4, 5 and 6, in all cases a greater number of extracted local features ($n$) improves the classification accuracy. Regarding to the local feature window sizes ($w$), while for no equalisation and global equalisation there is not a significant accuracy improvement, for the local equalisation method this parameter has an important influence on the error rate. The best results are obtained when local equalisation is used, more concretely a 0 % error rate is achieved for local equalisation with $w = 19$ and $n = 300$.

## 4. CONCLUSIONS

A system for palmprint recognition using local features has been presented. For this purpose we have used local features and we have compared 3 different preprocessing methods. The results for local equalisation are the best. On the other hand, in future works we will report our system performance by means of verification rates instead of classification rates.

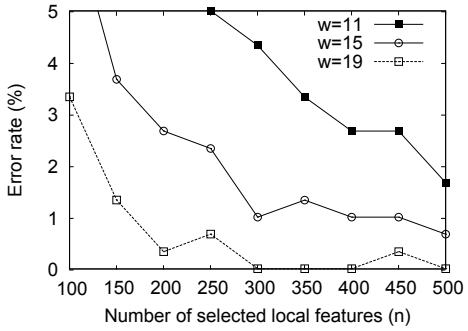



**Figure** 6: Results with Local Equalisation preprocessing.

## 5. REFERENCES


[1] PolyU Palmprint Database. http://www.comp.polyu.edu.hk/∼biometrics/.

[2] J. Daugman. The importance of being random: Statistical principles of iris recognition. *Pattern Recognition*, 36(2):279–291, 2003.

[3] Nicolae Duta, Anil K. Jain, and Kanti V. Mardia. Matching of palmprints. *Pattern Recognition Letters*, 23:477–485, 2002.

[4] S. Arya et al. An optimal algorithm for approximate nearest neighbor searching. *Journal of the ACM*, 45:891–923, 1998.

[5] Anil K. Jain and David Maltoni. *Handbook of Fingerprint Recognition*. Springer-Verlag New York, Inc, 2003.

[6] Anil K. Jain, Arun Ross, and Sharath Pankanti. A prototype hand geometry-based verification system. In *Proc. of 2nd Int'l Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), Washington D.C.*, pages 166–171, March 22-24 1999.

[7] D. Keysers, R. Paredes, H. Ney, and E. Vidal. Combination of tangent vectors and local representations for handwritten digit recognition. In *In SPR 2002, International Workshop on Statistical Pattern Recognition*, 2002.

[8] A. Kumar, D. C. M. Wong, H. C. Shen, and A. K. Jain. Personal verification using palmprint and hand geometry biometric. In *Proc. of 4th Int'l Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pages 668–678, Guildford (UK), June 9-11 2003.

[9] Ajay Kumar and David Zhang. Integrating shape and texture for hand verification. In *Proc. of Third International Conference on Image and Graphics (ICIG'04)*, pages 222–225, Hong Kong (China), 2004.

[10] Stan Z. Li and Anil K. Jain, editors. *Handbook of Face Recognition*. Springer, 2004.

[11] K Messer, J Kittler, M Sadeghi, A Kostin, and R. Paredes et al. Face authentication test on the banca database. In J.Kittler, M Petrou, and M Nixon, editors, *Proc. 17th Intern. Conf. on Pattern Recognition*, volume IV, pages 523–529, Los Alamitos, CA, USA, August 2004. IEEE Computer Society Press.

[12] R. Mohr, A. Picard, and C. Schmid. Bayesian decision versus voting for image retrieval. In *Proc of the CAIP-97*, 1997.

[13] R. Paredes, J. C. Pérez, A. Juan, and E. Vidal. Local Representations and a direct Voting Scheme for Face Recognition. In *Proc. of the Workshop on Pattern Recognition in Information Systems (PRIS 01)*, Setúbal (Portugal), July 2001.

[14] R. Paredes, E. Vidal, and F. Casacuberta. Local features for speaker recognition. *In SPR 2004.International Workshop on Statistical Pattern Recognition. LNCS 3138 of Lecture Notes in Computer Science*, pages 1087–1095, 2004.

[15] A. Ross, S. Dass, and A. K. Jain. A deformable model for fingerprint matching. *Pattern Recognition*, 38(1):95–103, 2005.

[16] M. Hatef R.P Duin J. Kittler and J. Matas. On combinig classifiers. *IEEE Trasn. on PAMI*, 1998.

[17] U. Uludag, A. Ross, and A. K. Jain. Biometric template selection and update: A case study in fingerprints. *Pattern Recognition*, 37(7):1533–1542, 2004.

[18] Jane You, Wenxin Li, and David Zhang. Hierarchical palmprint identification via multiple feature extraction. *Pattern Recognition*, 35:847–859, 2002.

[19] David Zhang, Wai-Kin Kong, Jane You, and Michael Wong. Online palmprint identification. *IEEE Transaction on Pattern Analysis and Machine Learning*, 25(9):1041–1050, September 2003.

# SALIENT POINTS DETECTION OF HUMAN FACES BY A CIRCULAR SYMMETRY INDEX BASED ON FISHER'S INFORMATION

*L. Capodiferro*[*]*, A.Laurenti*[**]*, P. Rava*[*]*, G. Jacovitti*[**]

[*]Fondazione Ugo Bordoni , Via B. Castiglione 59, 00142 Rome, Italy
Ph: +39 6 54802132; Fax: +39 6 54804401; email: licia@fub.it
[**]INFOCOM Dpt.,University of Rome "La Sapienza", via Eudossiana 18, 00184 Rome, Italy
Ph: +39 6 44585838; Fax: +39 6 4873300; email: gjacov@infocom.ing.uniroma1.it

## ABSTRACT

An image analysis scheme based on the measurement of Fisher information of local structures is presented. After a theoretical synthesis, a practical processing technique for calculating the Fisher information over the whole image, based on a filter bank is illustrated. Application of the method to facial recognition is outlined.

## 1. INTRODUCTION

Definition and detection of salient points is a basic process for image content analysis. Salient points considered in the present work are related to the Fisher information on orientation and size of small patterns in the observed image. More specifically, Fisher information maps of the patterns captured by a sliding window centered on every point of the image are first calculated. Salient points are then defined as the ones characterized by local maxima or minima of these maps and other associated maps. In particular, the Circular Symmetry (CS) map is especially suited for extracting salient points defined as centers of circular symmetric structures. In essence, the CS map is based on the fact that perfect circular patterns provide high Fisher's information on size and null information on orientation. By fact, the CS map is a useful index for measuring how much a pattern is circular, and it can be employed for detection and location purposes in many applications. In particular, we have employed Fisher and CS maps as novel tools for salient point detection in facial recognition applications. They are especially suited for detecting and locating pupils and corner like structures, such as lateral corners of eyes and mouth, as shown in the reported examples.
Since straightforward extraction of these features is computationally too expensive, they are indirectly calculated through a filter bank based on local image decomposition on a set of orthogonal, angular harmonic functions, constituted by the Gauss-Laguerre family. The outputs of this filter bank are finally combined with simple operations to give the Fisher information maps.
.

## 2. LOCAL FISHER INFORMATION ON ORIENTATION AND SIZE

Let $f(\mathbf{x})$ denote an observed image defined on the coordinates $\mathbf{x}=[x_1 \; x_2]^T$ of the real support $R^2$. Let us explore $f(\mathbf{x})$ with a sliding circularly symmetric window $w(\mathbf{x}-\boldsymbol{b})$ centered on the generic point $\boldsymbol{b}=[b_1 \; b_2]^T$ of the support. For each position $\boldsymbol{b}$ the window captures a portion of the image $w(\mathbf{x}-\boldsymbol{b})f(\mathbf{x})$. For mathematical convenience, let us refer to the relative coordinate system $\boldsymbol{\xi}=(\mathbf{x}-\boldsymbol{b})$ where $\boldsymbol{\xi}=[\xi_1 \; \xi_2]^T$ so that the captured pattern is:

$$g^w(\xi) = w(\xi)g(\xi) \qquad (1)$$

having posed $g(\boldsymbol{\xi})=f(\boldsymbol{\xi}+\boldsymbol{b})$.

In polar coordinates $r = \sqrt{\xi_1^2 + \xi_2^2}$ , $\gamma = \mathrm{tg}^{-1}\left(\frac{\xi_2}{\xi_1}\right)$ the captured pattern is :

$$g^w(r,\gamma) = w(r)g(r,\gamma) \qquad (2)$$

Our aim is to measure the Fisher information carried by any captured pattern with respect to the parameter pair constituted by its orientation and its scale. To this purpose, let us consider a situation where a copy of a captured pattern $g^w(r,\gamma)$ rotated by the angle $\varphi$, scaled by the factor $a$, and affected by additive noise is observed elsewhere. This actually occurs in many technical applications such image retrieval in databases, motion compensation in video sequences, robotics, etc.
It is expected that the Fisher information on these parameters, that we will refer to as orientation and scale "local Fisher information" (LFI) would constitute a useful tool for such applications. For instance, patterns having high LFI values with respect to orientation can be selected as salient points for performing reliable image rotation

estimates. Likewise, patterns having high LFI with respect to size can be selected as salient points for estimating zooming factors.

In [2] the calculus of the LFI has been conducted under the assumptions that:

- the additive noise is a white, zero-mean Gaussian Random Field with power density spectrum equal to $(N_0/4)$.

- the Fisher information carried by the multiplying window itself is negligible with respect to that of the windowed pattern. This is quite reasonable if the window is "smooth" enough (in practice we have adopted a gaussian shaped window).

It results [2] that the LFI on pattern orientation $\varphi$ can be written as follows:

$$J_\varphi = \frac{4}{N_0} a^2 \int_0^\infty r \left| w(ar) \right|^2 \int_0^{2\pi} \left[ \frac{\partial g(r,\gamma)}{\partial \gamma} \right]^2 d\gamma dr \qquad (3)$$

Likewise, it is shown in [2] that LFI on pattern scale $a$ can be written as

$$J_a = \frac{4}{N_0} \int_0^{2\pi} \int_0^\infty r^2 \left| w(ar) \right|^2 \left[ \frac{\partial g(r,\gamma)}{\partial r} \right]^2 r dr d\gamma \qquad (4)$$

## 3. CALCULUS OF THE LOCAL FISHER INFORMATION

The above expressions are not suitable for practical implementation, since their computation for every point ***b*** of the image support is cumbersome.

In order to derive an efficient processing scheme to calculate CS at each candidate point of the image, the pattern is expanded according to the following double indexed family of functions:

$$\begin{aligned} g^w(r,\gamma) &= e^{\frac{-\pi r^2}{\sigma}} g(r,\gamma) \\ &= \sum_n \sum_k \frac{1}{\sigma} C_{n,k} \mathcal{L}_k^{(n)} \left( \frac{r}{\sigma}, \gamma \right) \end{aligned} \qquad (5)$$

where the members

$$\begin{aligned} \mathcal{L}_k^{(n)}(r,\gamma) = (-1)^k 2^{(|n|+1)/2} \pi^{|n|/2} \left[ \frac{k!}{(|n|+k)!} \right]^{1/2} \\ \times r^{|n|} L_k^{(|n|)}(2\pi r^2)\, e^{-\pi r^2}\, e^{jn\gamma} \end{aligned} \qquad (6)$$

form the so called Gauss-Laguerre (GL) family and the parameter $\sigma$ serves to regulate the size of the sliding gaussian shaped window.

The GL functions are orthogonal and "harmonic angular". They possess many remarkable properties. Among others, they are obtained by differentiation of the gaussian function, are isomorphic with their Fourier transforms, and constitute admissible wavelets [3]. In equation (6), $n$ and $k$ denote the "angular" and the "radial" indices. Because of the orthogonality of these functions, the coefficients $C_{n,k}$ are calculated with the scalar product:

$$C_{n,k} = < e^{\frac{-\pi|\xi|^2}{\sigma}} g(\xi), \frac{1}{\sigma} \mathcal{L}_k^{(n)} \left( \frac{r(\xi)}{\sigma}, \gamma(\xi) \right) > \qquad (7)$$

Turning back to the whole image $f(\mathbf{x})$ the coefficients are calculated for every point **b** as:

$$\begin{aligned} C_{n,k}(b) &= < e^{\frac{-\pi|\xi|^2}{\sigma}} f(\xi + b), \frac{1}{\sigma} \mathcal{L}_k^{(n)} \left( \frac{r(\xi)}{\sigma}, \gamma(\xi) \right) >= \\ &< e^{\frac{-\pi|\xi|^2}{\sigma}} f(b - \xi), \frac{1}{\sigma} \mathcal{L}_k^{(n)} \left( \frac{r(-\xi)}{\sigma}, \gamma(-\xi) \right) >= \\ &f(\xi) * \left\{ e^{\frac{-\pi|\xi|^2}{\sigma}} \cdot \frac{1}{\sigma} \mathcal{L}_k^{(n)} \left( \frac{r(-\xi)}{\sigma}, \gamma(-\xi) \right) \right\} \Big|_{\xi = b} \end{aligned} \qquad (8)$$

In other words, the coefficients $C_{n,k}$ of the expansion (4), computed at each point ***b*** of the examined image, are obtained by convolving the observed image with the corresponding flipped members of the family (6), weighted with the window, i.e. by passing it through a bank of filters whose individual impulse responses are given by $e^{\frac{-\pi|\xi|^2}{\sigma}} \mathcal{L}_k^{(n)} \left[ r(-\xi), \gamma(-\xi) \right]$.

The next step is to calculate the LFI for each point ***b*** using the coefficients $C_{n,k}(b)$. In [2] it has been shown that

$$J_\varphi = \frac{4}{N_0} a^2 \sum_{n=0}^\infty \sum_{k=0}^\infty n^2 \left| C_{n,k} \right|^2$$

so that the $J_\varphi$ LFI map is obtained as:

$$J_\varphi(b) = \frac{4}{N_0} a^2 \sum_{n=0}^\infty \sum_{k=0}^\infty n^2 \left| C_{n,k}(b) \right|^2 \qquad (9)$$

As far as the information on scale is concernedit has been found in [1] that :

$$J_a = \frac{4}{N_0} \sum_{k=1}^\infty A(k) \left| C_{0,k} \right|^2$$

where the weighting factor *A(k)* versus the radial index *k* is plotted in Fig. 1.

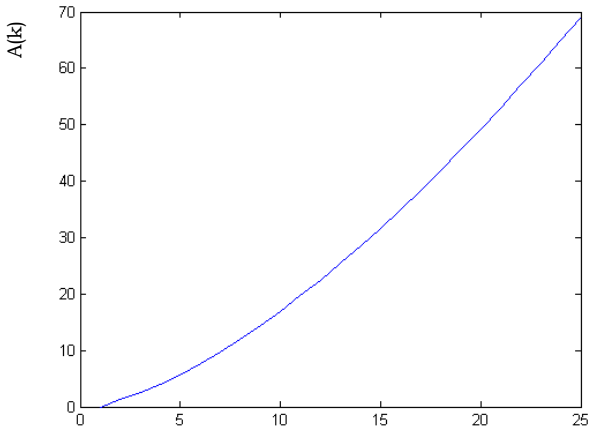



**Fig.1 Plot of *A(k)* versus the radial index *k***

so that the $J_a$ LFI map is obtained as:

$$J_a(b) = \frac{4}{N_0}\sum_{k=1}^{\infty} A(k)\left|C_{0,k}(b)\right|^2 \qquad (10)$$

Therefore, the LFI maps can be calculated directly from the coefficients of the Gauss-Laguerre expansion $C_{n,k}$ through (9) and (10) using a filter bank (for instance with FFT techniques). In practice, it suffices to truncate the Laguerre Gauss expansion to few terms (4x4) in order to obtain a good approximation of the Fisher's information. This truncation is done not only for the sake of computational convenience. In fact, in many operative scenarios it is useful to ignore small details, such as fine textures, cropping high-order Gauss-Laguerre components.

## 4. EXTRACTION OF CIRCULAR PATTERNS AND CORNERS

Equation (3) reveals that the LFI on orientation of a windowed pattern is proportional to the the centroid of the total weighted energy density of the azimuthal derivative along each radius. This means that $J_\varphi(b)$ is high in correspondence of vertices, and null in correspondence of centers of annular structures, which are invariant under rotation.

Equation (4) says that LFI on size is proportional to the integral sum over any orientation of the moment of inertia of the energy density of the radial derivative. This conversely implies that $J_a(b)$ is high in correspondence of the centers of annular structures and null in correspondence of vertices which are invariant to zooming.

This opposite behaviour of the considered LFIs does suggest to adopt the ratio:

$$\mathbf{CS=}\ \frac{J_a(b)}{J_\varphi(b)} \qquad (11)$$

as a Circular Symmetry indicator. In practice, the maxima of the CS map are employed for detecting and locating centers of patterns characterized by moderate azimuthal variations, irrespective of other specific attributes (radial profile) and, at some amount, irrespective of their size. The CS map has been successfully experimented for ball tracking in video sequences [1] where it proved quite robust with respect to non-idealities. Of course, the above observations do suggest that the inverse ratio $\frac{J_\varphi(b)}{J_a(b)}$ could be employed for extraction of vertices. However in the present work it has been experienced that such an indicator does not exhibit per se clear advantages with respect to the simpler classic Harris operator [4]. On the contrary, it has been noted that the $J_a(b)$ map is useful for improving the performance of the Harris operator. In fact, $J_a(b)$ presents local maxima in proximity of the vertices of corners, as expected. This property is helpful for eliminating false Harris operator detection events with a pruning stage.

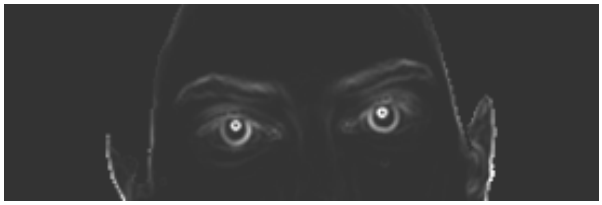
## 5. SALIENT POINT DETECTION IN HUMAN FACE RECOGNITION

These concepts have been applied to the field of facial recognition. In such applications, many different techniques of image digital processing and pattern recognition are jointly employed in order to achieve good detection performance. They range from shape and contour analysis to skin color discrimination, to geometrical and 3D modeling, etc. In this work we have tested the LFI indicators as additional tools for specific salient point extraction. In fact, human faces are characterized by well defined patterns pertaining to the class we have considered so far.

Let us first look at the problem of eyes detection. Figure 2 shows the maps of the Fisher' information components $J_a$ $J_\varphi$ and the CS= $\frac{J_a}{J_\varphi}$ index. This example clearly shows how the CS ratio is especially suited for extracting pupils, which are typical circular patterns.

*Eyes*

*Fisher's information on scale: Ja*

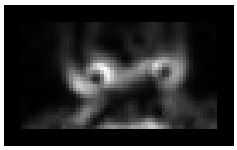


*Fisher's information on orientation: Jφ*

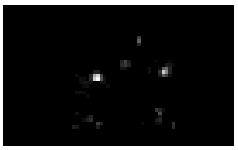
*map of the circular symmetry index CS=$J_a$ / $J_\varphi$*

**Fig.2 Example of eyes detection**

It has been experienced that the CS maps is also able to detect nostrils, which are interesting structures as well, as shown in the detail of fig. 3. This constitutes another useful geometric reference for facial recognition.



*Jφ*

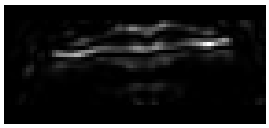


$J_a$ / $J_\varphi$

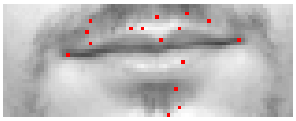
**Fig.3 Example of nostrils detection**

Finally, in fig. 4 the Harris and the $J_a$ maps are displayed side by side. It appears that many Harris map maxima do not correspond to corners. In fact, they are not accompanied by high values of $J_a$ in their immediate neighbours. Using the pruning criterion cited above, these maxima are ignored. This allows to eliminate many false detection points. Finally, the two extreme point are labeled as mouth corners.

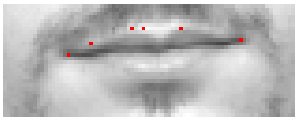
*Harris*



*$J_a$*



*Harris detection*



*$J_a$ based pruning*

**Fig.4 Example of mouth corners detection**

## 6. CONCLUSION

The LFI and CS maps presented here have been experimented on the FUB face database as tools for measuring interocular distance, eye width, nose position and mouth width, giving accurate results. These tools can be employed with many other techniques based on other features, such as skin color based segmentation [5] geometrical modeling [6] etc. to define sophisticated facial recognition systems. The LFI and CS maps are also useful for other biometric applications, such as pupil tracking and lip image segmentation.

## 7. REFERENCES


[1] L. Capodiferro, A. Laurenti, G. Monaco, M. Nibaldi, G. Jacovitti, "Tracking Of Circular Patterns In Video Sequences Based On Fisher's Information Analysis"

[2] Neri, G. Jacovitti, "Maximum Likelihood Localization of 2-D Patterns in the Gauss-LaguerreTransform Domain: Theoretic Framework and Preliminary Results" IEEE Trans. in Image Processing , Vol.13, No.1, pp. 72-86, January 2004.

[3] G. Jacovitti, A. Neri, "Multiresolution circular harmonic decomposition" IEEE Trans. on Signal Processing, Vol.48, No. 11, pp. 3242-3247, November 2000.

[4] C.Harris, M.Stephens, "A combined corner and edge detector", in *Proc. Jth Alvey Vision Conf.*, pp.147-151, 1988

[5] Peer, P. Kovac, J., AND Solina. "Human skin color clustering for face detection". EUROCON 2003- International Conference of Computer as a Tool, 2003.

[6] S.J.Jeng, H.M.Liao, Y.T.Liu, M.Y. Chern. "An Efficient Approach for Facial Feature Detection Using Geometrical Face Model", IEEE Proceedings of ICPR, 1996.

# A TWO-STAGE APPROACH TO THE DISCRIMINATIVE POINT SELECTION IN FACE IMAGES

*Daniel González-Jiménez*[1], *José Luis Alba-Castro*[1], *Enrique Argones-Rúa*[1*], *Josef Kittler*[2]

[1] Signal Theory and Communications Department, University of Vigo (Spain)
{danisub,jalba,eargones}@gts.tsc.uvigo.es
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH (UK)
J.Kittler@eim.surrey.ac.uk

## ABSTRACT

In this paper, we present a bi-layer method for the discriminative selection of key points in face images. In the rst level, locations are chosen from lines that depict facial structure, and which are thought to be inherently discriminative. In the second step, we select a subset of the initial constellation of nodes according to their classi cation accuracies. We report experimental results over the XM2VTS database on con gurations I and II.

## 1. INTRODUCTION

In automatic face recognition, selection of points for feature extraction is one of the most important steps in designing algorithms that encode local information. Several approaches have been advocated. In [1], the so-called ducial points are identi ed by minimizing a function which takes texture and geometrical distortion into account. Gabor lters are used both for point detection and feature extraction. The approach presented in [4] uses a rectangular grid atributed with morphological feature vectors in order to build a client model. Given a test image, the grid is deformed to match the training one, and pairwise comparisons between vectors are performed. Very good results are declared in [5]. The authors use a retinotopic sampling and computes special Gabor features to nd key points in face images, and use them for veri cation purposes. A quite different and simple system is presented in [7], where they look for points in faces where the intensity variance exceeds a threshold. Small patches surrounding such pixel are extracted and Principal Components Analysis is applied to get the - nal feature vectors. In this paper, we present a bi-layer node selection approach. First, a client-dependent facial structure is exploited to select possible discriminative locations and, in the second stage, a subset of these positions is kept based on their individual classi - cation accuracies. The hypothesis in both stages of discriminative point selection is that every user should keep the most representative features of himself and discard those shared with other users. The paper is organized as follows: Section 2 presents the shape-driven selection, feature extraction and matching stage, while section 3 explains the second layer of node selection. In section 4, we show our experimental results over the XM2VTS database [3], and a comparison with other methods reported in the literature. Finally, conclusions are drawn in section 5.

[*] Enrique Argones-Rúa performed the work while at the Centre for Vision, Speech and Signal Processing, University of Surrey

## 2. FIRST LAYER: SHAPE-DRIVEN POINT SELECTION AND MATCHING

In the rst step, a preliminary selection of points is accomplished through the use of shape information. Lines depicting face structure are extracted by means of a ridge and valley detector, and a set of locations $\mathcal{P} = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_n\}$ is chosen automatically by sampling from these lines. Figure 1 illustrates this procedure. These points are thought to be inherently discriminative, as their positions depend on individual face structure. Unlike other approaches, that try to look for universal features, i.e. eyes, tip of the nose, ..., our method nds points that are not con ned to belong to these universal characteristics. Furthermore, as this search just requires basic image operations, the selection of points in this step is rapid.

### 2.1. Feature extraction

A set of 40 Gabor lters $\{\psi_m\}_{m=1,2,\ldots,40}$, using the same con guration as in [1], are used to extract textural information from the region surrounding each shape-driven points. At point $\vec{p}_i = [x_i, y_i]^T$, we get the following feature vector:

$$\{\mathcal{J}_{\vec{p}_i}\}_m = \sum_x \sum_y I(x,y)\psi_m\left(x_i - x, y_i - y\right) \qquad (1)$$

where $\{\mathcal{J}_{\vec{p}_i}\}_m$ stands for the $m$-th coef cient of the feature vector extracted from $\vec{p}_i$. These vectors, called jets, are stored for further comparison. For a given face $I$ and its associated set of points, $\mathcal{P}$, we compute the set of responses $\mathcal{R} = \{\mathcal{J}_{\vec{p}_1}, \mathcal{J}_{\vec{p}_2}, \ldots, \mathcal{J}_{\vec{p}_n}\}$.

### 2.2. Point Matching

In a face authentication scenario, suppose we have a training image $I_{train}$ for client $C$. Given a test image $I_{test}$ claiming this identity, the system must decide if it is a true claim or, on the contrary,

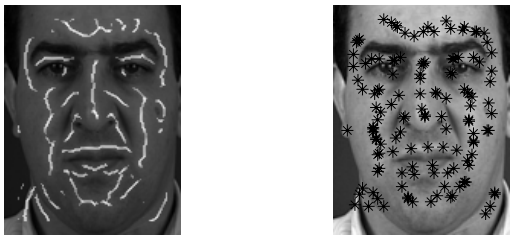


**Fig. 1**. Ridges superimposed over the original face image (left). Selection of points after rst stage (right)

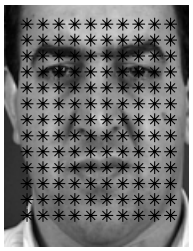


**Fig. 2**. Rectangular grid

an impostor attack. Let $\mathcal{P}_{train}$ and $\mathcal{P}_{test}$ be the set of points for $I_{train}$ and $I_{test}$ respectively. In order to compare feature vectors from both images, we decided to use a point matching algorithm based on shape contexts [2], so that we have a function $\xi$ that maps each point from $\mathcal{P}_{train}$ to a point within $\mathcal{P}_{test}$:

$$\xi\left(i\right):\vec{p_i}\Longrightarrow\vec{q}_{\xi(i)} \tag{2}$$

where $\vec{p_i}\in\mathcal{P}_{train}$ and $\vec{q}_{\xi(i)}\in\mathcal{P}_{test}$. Hence, the feature vector from the training image, $\mathcal{J}_{\vec{p}_i}$, will be compared to $\mathcal{J}_{\vec{q}_{\xi(i)}}$, extracted from $I_{test}$. The nal score between two images is:

$$\mathcal{S}=f_n\left\{<\mathcal{J}_{\vec{p}_i},\mathcal{J}_{\vec{q}_{\xi(i)}}>\right\}_{\vec{p}_i\in\mathcal{P}} \tag{3}$$

where $<\mathcal{J}_{\vec{p}_i},\mathcal{J}_{\vec{q}_{\xi(i)}}>$ represents the normalized dot product between correspondent jets, but taking into account that only the moduli of jet coef cients are used. In (3), $f_n$ stands for a generic combination rule of the $n$ dot products.

In section 4, we will present a comparison between the shape-driven selection and matching of points and a rectangular grid-based method, in which each node is located over the same facial region in every image, as it can be seen in gure 2.

## 3. SECOND LAYER: ACCURACY-BASED NODE SELECTION

In [4], the goal was to weigh the grid nodes according to their discriminatory power. They used a combination of statistical pattern recognition and support vector machines to identify which nodes were the most important for authentication purposes. Other approaches [9], [6] have selected and weighed the nodes from a rectangular grid (as the one shown in gure 2) based on a Linear Discriminant Analysis (LDA). This kind of analysis is possible due to the fact that a given node represents the same facial region in every image. In our case, we can not assume this, so we should use another method in order to select the most discriminative nodes.

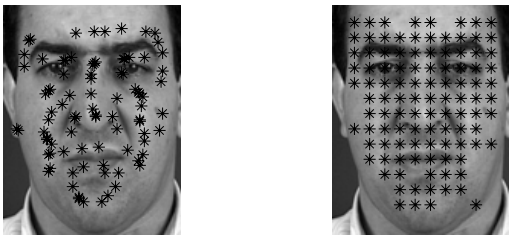


**Fig. 3**. Final set of points after layers 1 and 2 (left). Final set of points after layer 2 was applied to a rectangular grid (right)

The problem can be formulated as follows: given a training image for client $C$, say $I_{train}$, a set of images belonging to the same client $\left\{I_j^c\right\}$and a set of impostor images $\left\{I_j^{im}\right\}$, we want to nd which subset, $\hat{\mathcal{P}}\subset\mathcal{P}_{train}$, is the most discriminative. As long as each point $\vec{p_i}$ from $\mathcal{P}_{train}$ has a correspondent node in every other image (client or impostor, say $I_{test}$), we measure the individual classi cation accuracy of its associated jet $\mathcal{J}_{\vec{p_i}}$, and select the locations which achieve the best authentication rates, i.e., the ones with a Total Error Rate (TER) below a threshold $\tau$. Finally, only a subset of points, $\hat{\mathcal{P}}$, is chosen per image, and the score between $I_{train}$ and $I_{test}$ is given by:

$$\mathcal{S}=f_{\hat{n}}\left\{<\mathcal{J}_{\vec{p}_i},\mathcal{J}_{\vec{q}_{\xi(i)}}>\right\}_{\vec{p}_i\in\hat{\mathcal{P}}} \tag{4}$$

In section 4, we report the improvements due to this second layer selection, giving also a comparison between our method and other node selection-based approaches. Figure 3 (left) presents the set of points that were chosen after both layer selection for the image shown in gure 1.

## 4. DATABASE AND EXPERIMENTAL RESULTS

We tested our method using the XM2VTS database on both con gurations I and II of the Lausanne protocol [10]. The XM2VTS database contains synchronized image and speech data recorded on 295 subjects during four sessions taken at one month intervals. The database was divided into three sets: a training set, an evaluation set, and a test set. The training set was used to build client models, while the evaluation set was used to select the most discriminative nodes and to estimate thresholds. Finally, the test set was only used to measure the performance of the system. The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors, and 70 test impostors. Con gurations I and II of the Lausanne protocol differ in the distribution of client training and client evaluation data, representing con guration II the most realistic case. In con guration I, there are 3 training images per client and, in con guration II, 4 training images per client are available. Hence, for a given test image, we have 3 and 4 scores respectively, which can be fused in order to obtain better results.

## 4.1. Experiment 1

The rst test was intended to show the inherent discriminability of shape-driven points. We compared the performance of this rst layer node selection against a 10 by 13 rectangular grid, yielding an improvement in performance, as it is shown in table 1. We should remark that, for fair comparison between both methods, $13 \times 10$ shape-driven nodes per image were selected in this experiment. The median rule [8] (i.e. $f_n \equiv$ median) was used to combine the 130 dot products in both approaches. Also, the - nal 3 scores (con guration I) and 4 scores (con guration II) were fused using the median rule, leading to the nal score ready for veri cation.

## 4.2. Experiment 2

In the second test, we compared the full selection method (layers 1 and 2) against a downsampled rectangular grid, in which nal nodes were selected according to the accuracy-based method explained in section 3. The applied combination rules were the same as in 4.1. Again, as shown in table 2, our technique outperformed the rectangular grid approach. From tables 1 and 2, it is important to notice that the use of layer 2 also produced a clear improvement in the authentication rates for the rectangular grid. In gure 3 -right- we show the preserved points after accuracy-based node selection.

## 4.3. Experiment 3

Finally, our method was compared against a downsampled and weighted rectangular grid, whose nodes were chosen by means of a LDA-based selection method [9]. One of the main differences between both approaches is that the selected nodes by LDA are the same for every training image, while our algorithm chooses a different set of points per training image. Fusion techniques in our method were the same as in 4.1 (median-median). On the contrary, in [9] a different method for each con guration was used. In con- guration I, the fusion rule was MLP-based (MLP-MLP), while in con guration II, the combination of the selected node scores was done through LDA, and the nal 4 scores were fused using an MLP (hence, LDA-MLP). Figure 4 shows the selected nodes using this approach for both con gurations. The brightness of each node is proportional to its discriminative weight[1]. Finally, we must highlight that a client speci c threshold is used in the LDA-based method, whilst ours uses a different threshold per training image. Table 3 presents the results that demonstrate the better performance of the algorithm presented in this paper.

## 4.4. Results from other researchers

We have collected a set of results from two face competitions over the XM2VTS database, [11], [12], provided by different researchers. These results are presented in table 4. As we can see, this method is not the best performing algorithm over the database, although error rates are getting closer to others'. The retinotopic approach in [5], achieved a posteriori Equal Error Rate (EER) on the test set of 0.5%. Although this does not comply with the protocol exactly, it is still posible to compare performance. In table 4, TB and IDIAP (2000) used Gabor-based approaches.

[1] Recall that there are nodes with no weight - the ones that were discarded by the LDA selection

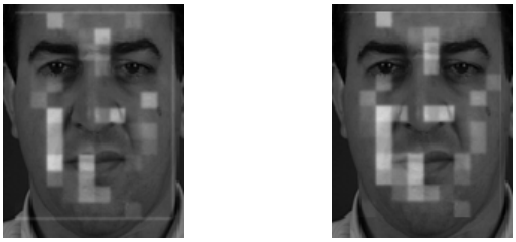


**Fig. 4**. LDA-based node selection: con guration I (left) and con- guration II (right)

| | Con guration I | Con guration II |
|---|---|---|
| | TER(%) | TER(%) |
| Rectangular | 8.08 | 5.65 |
| Ridges | **7.16** | **5.07** |

**Table 1**. Ridges (only layer 1 is used) vs. Rectangular grid

| | Con guration I | Con guration II |
|---|---|---|
| | TER(%) | TER(%) |
| Rectangular | 4.93 | 2.25 |
| Ridges | **3.61** | **2.09** |

**Table 2**. Ridges (layers 1 and 2) vs. Rectangular grid (layer 2)

| | Con guration I | Con guration II |
|---|---|---|
| | TER(%) | TER(%) |
| Rectangular | 4.45 | 3.80 |
| Ridges | **3.61** | **2.09** |

**Table 3**. Ridges (layers 1 and 2) vs. LDA-based node selection in a Rectangular grid (more details in [9])

| | Con guration I | Con guration II |
|---|---|---|
| | TER(%) | TER(%) |
| IDIAP (2000) | 16.6 | 15.0 |
| AUT (2000) | 14.2 | 9.7 |
| TB (2003) | 11.36 | 7.72 |
| IDIAP (2003) | 3.34 | 1.29 |
| UniS-NC (2003) | 1.48 | 0.75 |
| UPV (2003)*[7] | 3.48 | 2.30 |

**Table 4**. Results from other researchers [11], [12]. * is with full automatic registration.

## 5. CONCLUSIONS

In this paper, we have presented a method for selecting discriminative points in face images. A shape-driven stage is used to locate a set of points by exploiting individual face structure. Gabor jets are calculated at those positions and evaluated as individual classi ers so that nally , only the nodes which perform best are preserved. We have compared our technique against a rectangular grid-based approach and other researchers' methods, reporting good experimental results over the XM2VTS database. Furthermore, the computational load during the veri cation process is reduced, due to the smaller number of Gabor jets we need to compute from the selected points.

## 6. REFERENCES


[1] Wiskott, L., Fellous, J.M., Kruger, N., von der Malsburg, C.: Face recognition by Elastic Bunch Graph Matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 775-779, 1997

[2] Belongie, S., Malik, J., Puzicha J.,"Shape Matching and Object Recognition Using Shape Contexts," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 24, April 2002

[3] The XM2VTS database. http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/

[4] Tefas, A., Kotropoulos, C., Pitas, I., "Using Support Vector Machines to Enhance the Performance of Elastic Graph Matching for Frontal Face Authentication," IEEE Transactions on Pattern Analysis and Machine Intelligence Volume 23, Issue 7 (July 2001), pp. 735-746

[5] Smeraldi, F., Capdevielle, N., Bigun, J., "Face Authentication by retinotopic sampling of the Gabor decomposition and Support Vector Machines," in proc. Audio- and Video-based Person Authentication - AVBPA99

[6] Duc, B., Fischer, S., and Bigun, S., "Face authentication with sparse grid gabor information," In IEEE Proc. of ICASSP, volume 4, pp. 3053-3056, Munich 1997

[7] Paredes, R., Vidal, E., Casacuberta F., "Local Features for Biometrics-Based Recognition," in 2nd COST 275 Workshop. Biometrics on the Internet Fundamentals, Advances and Applications, March 2004, Vigo(Spain)

[8] Kittler, J., Hatef, M., Duin, R., and Matas, J., "On Combining Classi ers, " IEEE Transactions on Pattern Analysis and Machine Intelligence , Vol. 20, No. 3, March 1998

[9] Argones-Rúa, E., Kittler, J., Alba-Castro, J.L., González-Jiménez, D., "Information fusion for local Gabor features based frontal face veri cation, " submitted to International Conference on Biometrics 2006 (ICB 2006).

[10] Luttin, J. and Maˆtre, G. "Evaluation protocol for the extended M2VTS database (XM2VTSDB)." Technical report RR-21, IDIAP, 1998.

[11] Matas, J., et al. "Comparison of face veri cation results on the XM2VTS database" in Proceedings of the 15th ICPR, volume 4, pages 858-863, Los Alamitos, USA, September 2000. IEEE Computer Soc Press.

[12] Messer, K., et al. "Face Veri cation Competition on the XM2VTS Database", in Proc. AVBPA 2003, pp. 964-974

# 3D FACE RECONSTRUCTION FROM UNCALIBRATED IMAGE SETS

*Alexandros Moskofidis*[*], *Nikos Nikolaidis**

{amoskofi,nikolaid}@aiia.csd.auth.gr

[*]Department of Informatics, Aristotle University of Thessaloniki, GR-54124 Thessaloniki, Greece

## ABSTRACT

In this paper, we present a framework for reconstructing the 3D model of a human face from a set of uncalibrated facial images in different poses. The resulting 3D model of the face can be used in applications like face recognition and face verification. Our approach comprises two steps: in the first step we utilize a 3D reconstruction method to calculate the 3D feature coordinates of some salient feature points of the face, marked manually on the input images, whereas in the second step we use a mass springs finite elements method (FEM) to deform a generic face model, based on the cloud of points produced from the first step. We further enhance the resulting 3D model by projecting it into the input images and manually refining its node coordinates.

## 1. INTRODUCTION

The task of reconstructing an object in 3D space from its images (projections) is one of the most demanding in computer vision. In the past years the biggest attention was given to the calibrated reconstruction case (i.e. the case where the position of the camera relative to the object and the camera intrinsic parameters are known beforehand) whereas nowadays researchers try to tackle the uncalibrated reconstruction problem, where the input images are taken with a camera at random position and orientation with respect to the human face.

It is well known [1] that utilizing the epipolar geometry one can yield depth estimates for an object just from two images of it. Unfortunately, the obtained coordinates do not lie on the Euclidean space [2], which makes this representation not very useful. In order to upgrade the representation, extra information is required. This extra information can be obtained either from the camera position or from the camera intrinsic parameters. The latter can be calculated either from the use of special calibration patterns or from the images of our input set. The procedure of utilizing the images that we have in order to calculate the camera intrinsic parameters is called self calibration as opposed to calibration where some specific calibration patterns are used in order to calculate the camera calibration matrix.

There are numerous approaches to the uncalibrated 3D reconstruction problem in literature, the more characteristic of which are the work of Faugeras [3], Beardsley et al [4], Hartley [5] and Pollefeys [2], who wrote an excellent tutorial on the subject.

Our approach utilizes the 3D reconstruction algorithm presented by Pollefeys in [2] in order to calculate the 3D coordinates of some salient feature points of the face based on a small number of facial images where feature points are manually marked. We have chosen to use this approach because of its flexibility, due to the fact that the input images can be taken with an off the self camera placed at random positions. The intrinsic camera parameters can be calculated from the input image set.

We further incorporate a generic face model (the Candide face model) and deform it, using a finite element method (FEM), based on the point cloud obtained from the first step. On top of that, to further improve our resulting 3D model, we reproject it back to the initial images and fine tune it manually using an interface that was developed especially for this purpose. The resulting face model can be used along with the corresponding texture in biometric applications such as face recognition and face verification.

The rest of this paper is organized as follows. Part 2 describes in brief the first part of the proposed methodology, which is the 3D reconstruction of a set of salient features of the human face. In section 3 we describe the incorporation and the deformation of a generic head model (Candide head model) whereas in part 4 we provide some experimental results. In part 5 future directions are described and conclusions follow in part 6.

## 2. 3D RECONSTRUCTION

As already mentioned, we have used the algorithm proposed by Pollefeys in [2] in order to calculate the 3D coordinates of some salient features of the face. We will briefly explain the steps of algorithm for the sake of completeness of this paper. Readers interested in obtaining additional information can consult [2].

For our camera we have adopted the ideal pinhole – perspective camera model [6] where no radial distortion is present. In such a camera, the projection of an object point on an image plane is described by the following equation

$$\boldsymbol{m} = \boldsymbol{PM} \quad (1)$$

where $\boldsymbol{m} = [x, y, 1]^T$ are the point coordinates on the image plane, $\boldsymbol{P}$ is the 3x4 projection matrix and $\boldsymbol{M} = [X, Y, Z, 1]^T$ are the object point coordinates in 3D space. Note that we use the homogenous coordinates where the '=' sign indicates an equality up to an non-zero scale factor.

**Figure 1** : The input images

At the first step we manually select some salient feature points of the face in the input images and define their correspondences (figure 1) The coordinates of these feature points over the input images constitute the input to the 3D reconstruction algorithm. It has to be noted that we have used some easily recognizable and distinct feature points of the face such as the corners of the eyes, the corners of the mouth and the tip of the nose. Unfortunately it is very difficult to define a big number of feature points on the human face due to its lack of texture and characteristic points that can be uniquely identified over a number of images.

What comes next is the calculation of the Fundamental Matrix [1]. The calculation of the Fundamental Matrix is based on the first two images of the set. Those two images must be selected efficiently so that they correspond to viewpoints that are as far apart as possible but in the same time have all the feature points visible on both of them. The overall performance of the algorithm relies heavily on the efficient selection of these first two frames.

After the calculation of the Fundamental Matrix it is possible to obtain a reference frame which will eventually help us get an initial estimate of the depth for the selected feature points. Unfortunately, this representation does not lie in the metric space and thus additional procedures should be followed in order to upgrade it to metric.

Next the rest of the images of the input set are incorporated in the algorithm and the projection matrices that describe the projection of the face in each image of the set are evaluated.

In the subsequent step the algorithm performs an optimization which is based on all the images of the input set and thus refines the representation. This is called Bundle Adjustment [7] and it is the most computationally intensive part of the algorithm.

Finally the algorithm uses a self calibration technique in order to calculate the camera intrinsic parameters. These parameters are subsequently used to upgrade the representation to the metric space and yield the final cloud of points.

## 3. GENERIC MODEL DEFORMATION

The next part of the proposed approach deals with the incorporation of a generic face model, namely the Candide face model, into the reconstruction procedure.

The Candide face model has been developed by the Linkoping University [8] and in its current version has 104 nodes, distributed all around the human face and 184 triangles that connect those nodes creating a wire frame. The nodes of the model correspond to characteristic points of the human face e.g. nose tip, outline of the eyes, mouth etc. The feature points selected on the facial images are described in the previous section and should correspond to Candide nodes. A procedure for defining the correspondences between the 3D reconstruction of the selected feature points and the Candide model nodes was followed.

### 3.1 FEM Deformation

A mass spring finite element method was employed to deform the generic Candide model. The deformation process incorporates a list of pivotal points (our 3D reconstructed points from the first part of the algorithm), the Candide model and a list which contains the correspondences between the pivotal points and the Candide nodes, and produces a deformed model.

The FEM deformation can be outlined as follows: at first the Candide model undergoes global rotation translation and scaling so that it is roughly aligned with the cloud of 3D points. In order to determine the scale factor the mean distances between the two corners of the eyes and the two corners of the mouth were evaluated both in the point cloud and the Candide model and their ratio was used as the scale factor. Then the model was translated so that the center of mass of the point cloud coincides with the center of mass of the corresponding model nodes.

Furthermore the Candide model has to be appropriately rotated. To achieve this, a triangle whose vertices are the outer tips of both eyes and the tip of the nose was defined. The same triangle was defined for the corresponding nodes of the Candide model and the model was rotated so that the outwards pointing normal vectors of the two triangles are aligned. The deformation process moves the corresponding nodes of the Candide model so that they coincide with points of the cloud and deforms the rest of the nodes. As it is obvious from the latter, pivotal points must spawn the entire face, otherwise the deformation process will produce poor results.

## 3.2 Manual Refinement

After the application of the deformation we obtain a model that fits the individual's face depicted in the input set of images. Unfortunately, due to limitations on the 3D reconstruction algorithm, the deformation process and to errors in the selection of the feature points coordinates, the output model may not be ideal, in the sense that some nodes may not have the correct position in 3D space. Therefore a manual refinement procedure is adopted.

According to this procedure, we reproject the deformed face model in every image of the input set and manually change the location of certain model nodes in the 2D domain. In order to return to the 3D domain from the manually refined projections, a triangulation process is used [6]. This was facilitated from the fact that the projection matrices for each frame are available from the 3D reconstruction algorithm.

In order to be able to use the triangulation method to estimate the 3D coordinates of a model's node we must specify manually the new positions of the nodes in two frames. By so doing we can yield new, improved coordinates, in the 3D space. When the manual adjustment of the selected nodes is finished the deformation process is applied once again but this time with an extended set of pivotal points – the initial cloud of points produced from the 3D reconstruction algorithm along with the additional 3D coordinates of the points that have been manually refined.

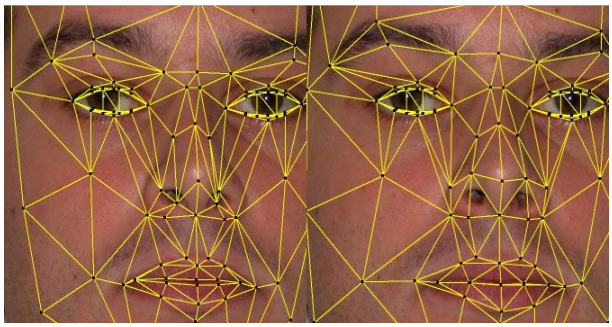


**Figure 2** : The manual refinement procedure

The manual refinement procedure is presented in figure 2 which depicts the projection of the deformed model into an image of the input set prior and after the manual refinement. It is evident that with the manual refinement the generic model can fit more efficiently to the individual's face.

## 4. EXPERIMENTAL RESULTS

For our experiments we have used a minimal set of 3 images of human faces in different positions and we have selected and matched manually the feature points across those images.

Feature points were selected on the two corners of the mouth, the two corners of the eyes and on the tip of the nose, as shown on figure 1.

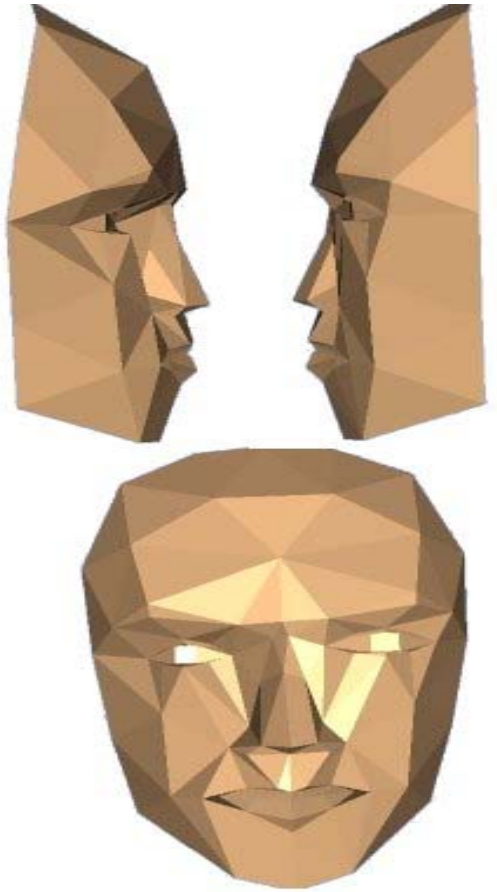


**Figure 3** : The deformed Candide Model

The deformed Candide model derived from the facial images of figure 1, after applying limited manual refinement is presented in figure 3. Obviously the deformed model is not perfect which can be attributed to the errors in the feature selection process as well as to the limited resolution of the Candide model.

The performance of the 3D reconstruction algorithm was evaluated based on the reprojection error namely the Euclidean distance between the manually selected feature points and the projections of the derived 3D features. Results for the model presented in figure 3 can be seen in table 1.

| Manually Selected Coordinates | Calculated Coordinates | Reprojection Error (pixels) |
|---|---|---|
| (1131,1151) | (1131,1151) | (0,0) |
| (1420,1164) | (1420,1164) | (0,0) |
| (1050,776) | (1051,775) | (-1,1) |
| (1221,786) | (1218,788) | (3,-2) |
| (1392,794) | (1395,795) | (-3,-1) |
| (1567,793) | (1566,792) | (1,1) |
| (1244,957) | (1244,957) | (0,0) |

**Table 1** : Reprojection error – image 3

One can observe that the reprojection error is very small and does not exceed 3 pixels for input images of dimensions 2560x1920. Similar results were obtained when the algorithm was applied to other image sets.

A number of provisions can be taken in order to make the algorithm more robust. The camera positions used for capturing the image should be sufficiently apart but at the same time care has to be taken in order to ensure that all feature points are visible in the first three images of the set. The most error prone part of the algorithm is the initial triangulation (depth estimation) where a small angle between the viewpoints used to acquire the two images can have a severe effect on the overall reconstruction.

Moreover through experimentation we have reached the conclusion that the quality of the results is mainly affected by the quality of the input features i.e. whether corresponding points selected on the images are indeed projections of the same points on the 3D space. Thus care should be taken in order to select these points as accurately as possible.

## 5. FUTURE WORK

Our work in the field of modeling a human face from a set of uncalibrated images is not complete yet. In the future we plan to incorporate some new techniques that will aim towards a more robust 3D reconstruction, namely a new method for bundle adjustment that besides the reprojection error will incorporate additional constraints derived from the geometry of the human face (e.g. the relative distances of the eyes, mouth etc).

Furthermore, we are experimenting with different techniques for the deformation of the generic Candide face model to be used as an alternative to the finite elements method.

## 6. CONCLUSIONS

In this paper we have presented a framework for the challenging task of reconstructing a face in three dimensions from a set of uncalibrated images.

In the first part of the proposed approach we used a 3D reconstruction algorithm proposed by Pollefeys [2] to calculate the 3D coordinates of some salient facial feature points manually selected on a set of images. At the second part of the algorithm, we use a generic face model in order to produce a more detailed representation of the object's face. This is substantially an interpolation process with all the advantages and disadvantages that this entails. In order to obtain a more detailed and accurate model of the individual's face an iterative manual refinement process is employed, which can improve the quality of the resulting face model. The experimental results prove that the proposed methodology can yield very satisfactory 3D reconstructions.

## 7. REFERENCES


1. Andrew Zisserman, Richard Hartley, "Multiple View Geometry in Computer Vision", Second Edition, Cambridge University Press 2004
2. Marc Pollefeys, "Tutorial on 3D Modelling from Figures", June 2000 (http://www.esat.kuleuven.ac.be/~pollefey/tutorial/)
3. Olivier Faugeras, "What can be seen in three dimensions with an uncalibrated stereo rig", ECCV '92, p. 563-578
4. Paul A. Beardsley, Andrew Zisserman, D.W. Murray, "Sequential Updating of Projective and Affine Structure from Motion", International Journal of Computer Vision, June/July 1997, p. 235-259
5. Richard Hartley, "Euclidean Reconstruction from Uncalibrated Views", Second joint European – US Workshop on Applications of Invariance in Computer Vision, 1993 p. 237-256
6. Emanuele Trucco, Alessandro Verri, "Introductory Techniques in 3-D Computer Vision" , Prentice Hall 1998
7. Bill Triggs, Philip McLauchlan, Richard Hartley, Andrew Fitzgibbon, "Bundle Adjustment – A Modern Synthesis", International Workshop on Vision Algorithms 1999 p.298-372
8. J. Ahlberg, CANDIDE-3 -- an updated parameterized face, Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linköping University, Sweden, 2001.



The authors would like to thank George Moschos for his implementation of the Finite Elements Method library.

# COMPARISON OF ILLUMINATION NORMALIZATION METHODS FOR FACE RECOGNITION*

*Mauricio Villegas Santamaría and Roberto Paredes Palacios*

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022 Valencia (Spain)
{mvillegas,rparedes}@iti.upv.es

## ABSTRACT

Illumination invariance is one of the most difficult properties to achieve in a face recognition system. Illumination normalization is a way to solve this problem. In this paper we compare several illumination normalization techniques, including traditional and proposed ones. An error rate less than 1% is achieved using the Yale Face Database B.

## 1. INTRODUCTION

The design of an automatic face recognition system has several challenges, being one of these the invariance to lighting changes. In an face image, changes due to illumination are generally greater than the differences between individuals. This makes it necessary to take into account the illumination in order to make the system reliable in less constrained situations.

Many methods have been proposed [1] [5] [6] [8] [10] [11] to compensate illumination variations for face recognition, which are based on different ideas. One way to make face recognition invariant to illumination changes is to process the images prior to the the classifier stage. This process, called illumination normalization, attempts to transform an image with an arbitrary illumination condition to a standard illumination invariant one. In this paper we present the results of the experiments made comparing different illumination normalization techniques.

The paper is organized as follows. Section 2 briefly describes the different techniques for illumination normalization that are compared in the experiments. First, the global normalization methods are described, which are more commonly used in digital image processing, followed by three proposed local normalization methods. In section 3 the results of the experiments are presented, analyzed and compared with other methods reported in the literature. The final section includes the conclusions and the directions for future research.

*Work supported by the "Agencia Valenciana de Ciencia y Tecnología (AVCiT)" under grant GRUPOS03/031 and the Spanish Project DPI2004-08279-C02-02

## 2. NORMALIZATION METHODS

### 2.1. Global normalization methods

In digital image processing when an image is not well illuminated, there are several methods to correct the problem. If the image must keep the reality appearance, these methods are normally applied to the image globally. The following are the global methods used in the experiments.

#### Gamma Intensity Correction (GIC)

The gamma transform of an image is a pixel transform in which the output and input are related by exponentiation

$$f(I(x,y)) = I(x,y)^{1/\gamma} \quad (1)$$

Depending on the value of $\gamma$ the output image is darker or brighter. In GIC the image is gamma transformed as to best match a canonically illuminated image $I_C(x,y)$. To find the value of $\gamma$ the following equation must be solved

$$\gamma = \arg\min_{\gamma^*} \sum_{x,y} [I(x,y)^{1/\gamma^*} - I_C(x,y)]^2 \quad (2)$$

In our implementation the value of $\gamma$ is approximated using the golden section search [1].

#### Histogram Equalization (HE)

In histogram equalization the result is obtained using the cumulative density function of the image as a transfer function. The result of this process is that the histogram

becomes approximately constant for all gray values. For an image of size $M \times N$ with $G$ gray levels and cumulative histogram $H(g)$ this transfer function is

$$T(g) = \frac{G-1}{MN} H(g) \tag{3}$$

### Histogram Matching (HM)

Histogram matching, also known as histogram fitting or histogram specification, is the generalization of HE. The objective is to have a resulting image with a histogram similar to a specified one. For illumination normalization the image is histogram matched taking as reference the histogram of a well illuminated image.

Analog to histogram equalization, we must find a transfer function. This can be done by first calculating for the histogram of the image and the specified one ($H(g)$ and $S(g)$) the transfer function that maps them to uniform histograms ($T_{HU}(g)$ and $T_{SU}(g)$) using equation 3. Then the transfer function for the specified histogram is inverted ($T_{US}(g)$) and the final transfer function is the following [4]

$$T(g) = T_{US}(T_{HU}(g)) \tag{4}$$

### Normal Distribution (NORM)

This technique normalizes the image by assuming the gray values form a normal distribution. The idea is to make the mean ($\mu_r$) and the standard deviation ($\sigma_r$) of the resulting image to zero and one respectively. For an image of mean $\mu_i$ and standard deviation $\sigma_i$, the output image is calculated using the following equation.

$$f(I(x,y)) = \frac{I(x,y) - \mu_i}{\sigma_i} \tag{5}$$

## 2.2. Local normalization methods

The local normalization methods have the disadvantage that the output is not necessarily realistic. In the problem at hand the objective is not to have a realistic image but to obtain a representation of the face that is invariant to illumination, while keeping the information necessary to allow a discriminative recognition of the subjects. With this idea in mind, it makes sense to use local illumination normalization methods this type of application.

The local methods used are: Local Histogram Equalization (LHE), Local Histogram Matching (LHM) and Local Normal Distribution (LNORM). They are the same as their global counterparts but applied locally. By applying a function locally we mean the following. We crop the image, starting in the up left corner, with a window size considerably smaller than the image size. The global normalization function is applied to the cropped image. This process is repeated by moving the crop window pixel by pixel all over the image and for each one applying the normalization function. Because the crop windows overlap, the final pixel value is the average of all the results for that particular pixel.

## 3. EXPERIMENTS AND RESULTS

For the experiments, we used cropped [7] faces from the Yale Face Database B [2] and the extended Yale Face Database B [3]. This database includes images of 38 people under 64 different illumination conditions but only for the frontal pose. These images are available on the Internet, originally they have a size of 168×192 but we resized them to 75×85 pixels.

Three of the methods require a model of a well illuminated face, they are GIC, HM and LHM. For this model we used the average of the cropped extended Yale Face Database B followed by contrast enhancement. Because the different illumination conditions are symmetrical, the average is a well illuminated face.

All the images of each person where taken in a period of two seconds, the only variation is illumination and there is practically no change in expression. So for recognition, a simple distance classifier can be used. For the experiments we chose the nearest neighbor classifier using the Euclidean distance between the images.

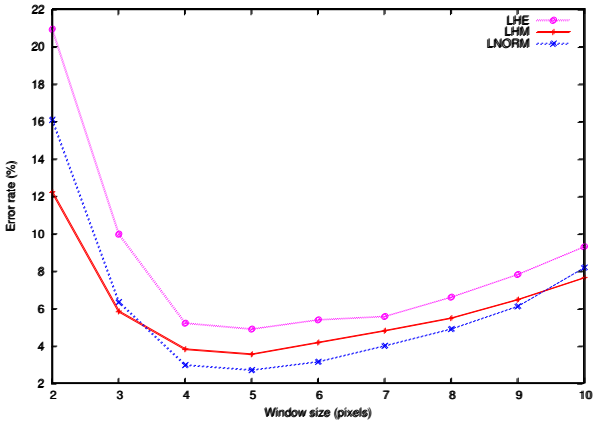



Figure 1: Classification error for the extended Yale Face Database B using local illumination normalization methods varying the window size.

## 3.1. First experiment

The first experiment was to calculate the optimal window size for each of the local normalization methods. This was done using the extended database (38 people). For each person, 5 well illuminated images were used for training and the remaining 59 for test. The classification error was calculated for the global methods and for the local methods using different window sizes. The results are summarized in figure 1 and table 1. Some example

| Normalization Method | Error rate (%) |
|---|---|
| NO NORMALIZATION | 70.86 |
| NORM | 53.71 |
| GIC | 48.90 |
| HISTEQ | 48.00 |
| HMATCH | 43.02 |
| LHE 5×5 | 4.89 |
| LHM 5×5 | 3.54 |
| LNORM 5×5 | 2.69 |

Table 1: Classification error for the extended Yale Face Database B for local, global and no illumination normalization.

images with no processing and normalized with each of the methods can be found in figure 2.
The global normalization methods improve the classification error, being HMATCH the best one. But all the local methods have a classification error an order of magnitude lower, as was expected. The best method is LNORM, moreover one of the advantages of LNORM is that it does not depend on a face model, so it is expected to work well when there are changes of expression and pose. Also, for the optimal window size, it is the fastest of the local methods.

## 3.2. Second experiment

The second experiment was to calculate the classification error for the non extended Yale Face Database[1] (10 people, 640 images in total) using the best normalization method (LNORM with a window size of 5×5). The purpose of this is to be able to compare with results found in the literature. It is custom to divide the images in subsets depending on the angle of the light source with respect to the camera axis. The subsets are: S1 is for angles lower than 12° (70 images), S2 for 12°-25° (120 images), S3 for 25°-50° (120 images), S4 for 50°-77° (140 images) and S5 above 77° (190 images). So the subsets are ordered according to the quality of the illumination, being S1 the best and S5 the worst. Refer to [6] for more information about this arrangement.

The results are summarized in table 2. For each subset we calculated five classification errors. The five classification errors differ on the training set, in each case these were selected from a different subset. In all the cases we took 5 images per subject for training, 50 in total. After the random selection the remaining images are used for test. To have more reliable results, the classification error is the average using ten different randomization seeds for the selection of the training set.

Table 2 shows that when training with well illuminated

[1]The four corrupt images and the three images with no light were not removed from the database

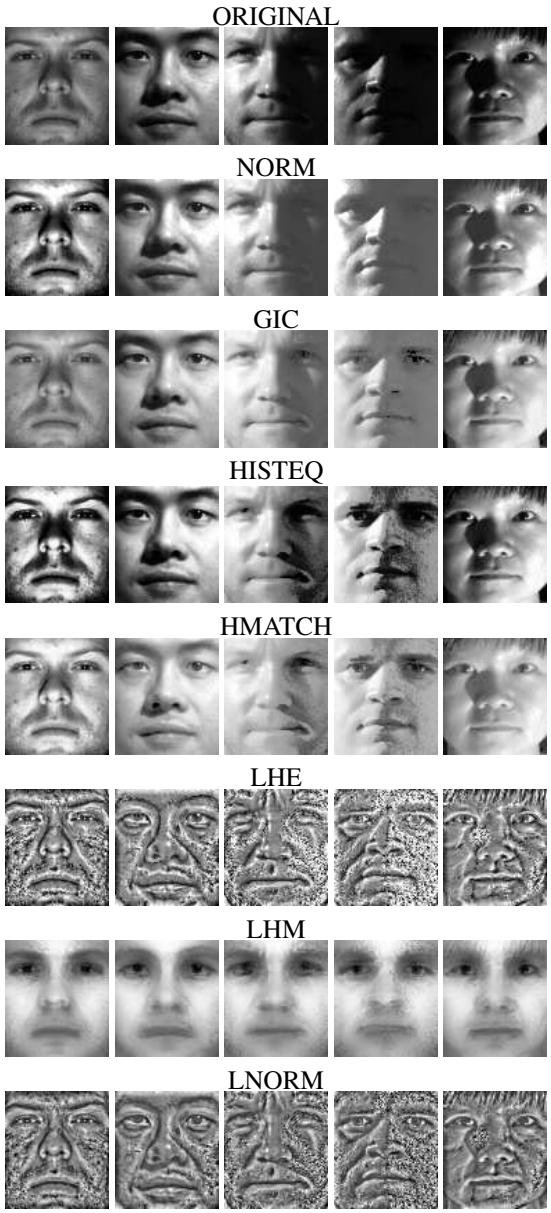



Figure 2: Example images from the Yale Face Database B and the normalized versions from the experiments. The normalized methods are sorted according to the classification error.

| | S1 | S2 | S3 | S4 | S5 | ALL |
|---|---|---|---|---|---|---|
| Tr S1 | 0.00 | 0.00 | 0.00 | 1.36 | 1.68 | 0.86 |
| Tr S2 | 0.00 | 0.00 | 0.58 | 1.00 | 4.26 | 1.72 |
| Tr S3 | 0.00 | 0.83 | 0.86 | 1.21 | 0.05 | 0.57 |
| Tr S4 | 5.57 | 1.92 | 1.92 | 3.33 | 0.32 | 2.05 |
| Tr S5 | 10.14 | 8.67 | 4.67 | 1.57 | 0.00 | 4.29 |

Table 2: Classification error for each subset of the Yale Face Database B using LNORM 5×5. The rows represent the result when training with images from each subset, 5 randomly selected images for each subject.

images (Tr S1) the global error rate is low, having the highest error for the extreme illumination. If we train with worse illumination (Tr S5) it performs better for the extreme illumination but in general error rates get higher. It is interesting to note that in average the lowest error rate is when training with S3. We can say from this that when an application has to overcome a large variability in illumination, it is advisable to include in the training set some images with not ideal illumination.

In the paper from Shan [10] they compare several normalization methods, LNORM outperforms all of them. In particular, Quotient Illumination Relighting (QIR), achieves 9.4% and 17.5% for S4 and S5 compared to 1.36% and 1.68% with LNORM when training with S1.

In [6] the Nine Points of Light (9PL) method is proposed and compared with other six methods all of which need subsets 1 and 2 for training. Although 9PL is supposed to perform very well for S4, the error rate is 2.8% slightly higher than LNORM for the equivalent test training with S4. The only method that outperforms LNORM is conescast that has an error rate of zero for S1-S4, for S5 the error was not calculated. For this method the error rate is lower, but it needs a lot more images for training.

The last method we are going to compare was presented in [11]. For S3 and S4 the error rates are 0.3% and 3.1%, compared to 0.00% and 1.36% of LNORM. This method has the advantage that only needs one image for training, but the error rates are higher and the technique is a lot more complex so it could be impractical for many applications.

## 4. CONCLUSIONS

It is possible to develop illumination normalization methods for face recognition that are a simpler than most of the ones found in the literature, having comparable or even better recognition results. This was proved with the Local Normal Distribution (LNORM) method presented in this paper. Although this normalization method produces a non realistic image that apparently has a lot of noise, the results indicate that it's a good illumination invariant representation for face recognition.

The method was only tested for images with the same pose and expression. In our future work we plan to make tests with images that have these types of changes. This will make it possible to compare our results with other methods like [5]. For those tests we will be using local features with the k nearest neighbors classifier [9].

## References


[1] O. Arandjelović and R. Cipolla. An illumination invariant face recognition system for access control using video. *In Proc. British Machine Vision Conference*, September 2004.

[2] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. In *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pages 277–284, 2000.

[3] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

[4] R. Gonzalez and R. Woods. *Digital Image Processing*. Addison-Wesley Publishing Company, 1992.

[5] R. Gross and V. Brajovic. An image preprocessing algorithm for illumination invariant face recognition. In *4th Int. Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*. Springer, June 2003.

[6] K.C. Lee, J. Ho, and D. Kriegman. Nine points of light: Acquiring subspaces for face recognition under variable lighting. In *CVPR*, volume 1, pages 519–526. IEEE Computer Society, 2001.

[7] K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[8] Stan Z. Li and Anil K. Jain. *Handbook of Face Recognition*. Springer, 2005.

[9] R. Paredes, E. Vidal, and F. Casacuberta. Local features for biometrics-based recognition. In *2nd COST 275 Workshop. Biometrics on the Internet Fundamentals, Advances and Applications*, 2004.

[10] B. Cao S. Shan, W. Gao and D. Zhao. Illumination normalization for robust face recognition against varying lighting conditions. In *International Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, held in conjunction with IEEE ICCV-2003, pages 157–164, October 2003.

[11] L. Zhang and D. Samaras. Face recognition under variable lighting using harmonic image exemplars. In *CVPR*, volume 1, pages 19–25. IEEE Computer Society, 2003.

# SPEAKER RECOGNITION

# METHODOLOGY OF SPEAKER RECOGNITION TESTS IN SEMI-REAL VOIP CONDITIONS

*Piotr Staroniewicz, Wojciech Majewski*

Institute of Telecommunications, Teleinformatics and Acoustics
Wroclaw University of Technology

## ABSTRACT

The paper presents methods of signal degradation in packet-based voice transmission. The decisive factors for that kind of transmission are briefly reviewed. Apart from the fully simulated VoIP conditions which are most frequently applied in tests also more closer to the real VoIP transmission models are proposed. Several tools available nowadays let the experiments be carried out in highly controlled network parameters on the one hand and semi-real conditions on the other.
The preliminary test results of speaker identifications in the three network conditions together with quality assessments of the above are presented.

## 1. INTRODUCTION

Data is carried most efficiently on packet networks and on the other hand integration of voice and data on a single network offers significantly improved efficiency. In modern networks data became the major type of telecom traffic which leads to using packet–based transmission also in the integrated networks. Packet-based transmission of voice has important implications for voice quality [16]. Voice quality in packet transmission depends on the following factors:

1. The speech codecs
2. The delay across the network and variation in the delay (jitter)
3. The packet loss across the channel
4. The echo.

Most nowadays experiments carried on speech signal (including speech and speaker recognition and speech quality assessment) use fully simulated simple network model. Simulating the network transmission enables the full control of the experiment conditions but does not guarantee the full reliability of the obtained results. On the other hand, carrying out the experiments in the real conditions comes across several obstacles, i.e difficulties with controlling the transmission parameters or a problem of enabling the repetition of a certain experiment in the same network conditions. On the second COST275 workshop in Vigo the conception of creating the real conditions VoIP database was presented [7]. In this paper the proposed system was set up and preliminary experiments were carried out.

### 1.1 Speech codecs in packet transmission

Selecting the codec is an essential problem for speech transmission. A codec converts the analog voice signal to a digitized bit stream at one end of the channel and returns it to its analog state at the other.

| Codec | Type | Bit rate | Frame size | Total delay |
|---|---|---|---|---|
| G.711 | PCM | 64 kbps | Depends on packet size | |
| G.726 | ADPCM | 32 kbps | | |
| G.729 | CS-ACELP | 8 kbps | 10 ms | 25 ms |
| G.729A | CS-ACELP | 8 kbps | 10 ms | 25 ms |
| G.723.1 | MP-MLQ | 6.3/5.3 kbps | 30 ms | 67.5 ms |
| GSM.EFR | ACELP | 12.2 kbps | 20 ms | 40 ms |

**Table 1:** Characteristics of speech codecs used in packet networks.

Table 1 shows typical voice over IP codecs [7,10]. The G.711 codec provides high quality of the connection with the PCM (pulse code modulation) coding. It is a waveform codec which operates at 64 kbps and which packet size is set arbitrary (for 20ms packetization the delay is 20ms). The G.726 codec is also the waveform codec which also has the packet size set arbitrarily. It reduces the data rate (degrading the quality) and uses ADPCM (adaptive differential pulse code modulation). For both above codecs processing delay is negligible and main delay associated with the use of them is the packetization delay. This is equivalent to the packet length which is usually from 10 to 40 ms.

The CELP (code excited linear predictive) codecs are based on a model of the acoustics of the vocal tract during the speech production which makes the transmission with a lower data rate possible (typically from 4 to 16 for telephony applications). Therefore CELP codecs create more delay than waveform codecs. The G.729 is the 8 kbps codec with good delay characteristics (due to short frame) and acceptable voice quality. The G.729A has a reduced coding complexity and identical decoding with the equivalent voice quality in comparison to the above. The G.723.1 codec based on multi-pulse maximum likelihood quantization is applied in the bandwidth limited transmission channels. The GSM.EFR is a wireless codec which uses 20 ms frame length.

## 1.2 Transmission delay and packet loss

Beside speech coding, the quality of VoIP is determined mainly by packet loss and delay. If packet is lost the quality degrades and on the other hand, if the packet delay is too high and misses the playout buffer, it leads to a late loss. If packet is lost or has a large delay, the next one is also likely to do so.

The end-to-end packet delay, also known as a latency, includes the time taken to encode the sound as a digital signal, the signal's journey through the network and the regeneration of it as a sound at the receiving end. Descriptions of the components contributing to the end-to-end delay are presented in Table 2. In the IP network packets travel independently and they are interspersed with packets from other network traffic along the way. There are two ways of packet loss. First, they can be lost at network nodes because of an over-flow in the buffer or because a congested router discards them. Second, packets can be delayed if they take a longer route causing that they can arrive after the prescribed delay and lose their turn.

| Delay sources | Ranges | Description |
|---|---|---|
| **Transmission Delays** | 1-100 ms for terrestrial; ~300 ms for geostationary satellite | from short local propagation delays to longest around globe |
| **Sender Delay** | | |
| Codec | 2-100 ms | Includes encoding and packetization delay, for single IP hop, one frame per packet |
| Other DSP | 0-30 ms | PLC, noise suppression, silence suppression, echo cancellation |
| **Receiver Delays** | | |
| Delay for jitter buffer | 1-20 ms | depends on utilization and whether congestion control is used |
| Multiple frames per packet | 10-60 ms | time of additional frames beyond one |
| Interleaving | 5-90 ms | depends on size of frames and packets |

**Table 2**: Types and causes of delays.

# 2 FROM SIMULATED TO REAL VOIP CONDITIONS

## 2.1 Simulation with Gilbert models

Studies on the distribution of the packet loss on the Internet [4,6,8] have concluded that this process could be approximated by Markov models [2,5]. The two states Markov model, also known as the Gilbert model (Fig.1) is used most often to capture the temporal loss dependency. In Fig.1, $p$ is probability that the next packet is lost, provided the previous one has arrived, $q$ is the opposite and 1-$q$ is the conditional loss probability. A more general $n$th-order Markov chain can also be used for capturing dependencies among events. The next event is assumed to be dependent on the last $n$ events, so it needs $2^n$ states. Usually it is enough to use up to six states but sometimes it can be 20 to 40. In the Markov model all the past n events can affect the future whereas in the extended Gilbert (the Gilbert model is a special case of the extended Gilbert model when $n$=2) model only the past $n$ consecutive loss events can do. That is why it does not fully capture the burstiness or clustering between loss and inter-loss distance metric. ILD (inter-loss distance metric) can be used to prevent it.

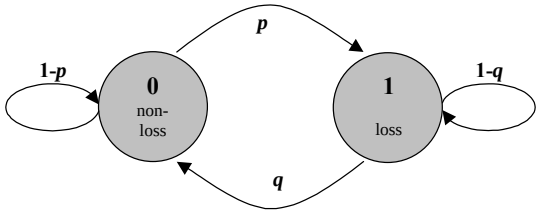



**Figure 1**: Gilbert model

## 2.2 Semi-real conditions

Schemes of possible methods of network transmission degradation were presented in Fig.2. The only possible way of full control of the network conditions between two VoIP sets is a tool realizing packet impairment generation. This process can be realized by several tools (e.g.: *NetDisturb* [19], *ISI Network simulator* [9], *Shunra Cloud* [14] etc.) on a PC equipped with two LAN interfaces. Generation of impairment (packet loss or delay) can be done with the use of certain mathematical rules (i.e. Gilbert model, extended Gilbert model etc.). In this case however we resign from the test made in the real network structure but we can use real sending and receiving VoIP devices (including codecs) in real acoustic conditions. The other way is generating the IP traffic in a certain network structure (Fig.2). The traffic generator (e.g.: *IP Traffic*[17], *LanTraffic*[18]) should be a system which is able to manage several IP connections simultaneously. Usually this kind of tool is composed of two parts: the sender and the receiver. The sender part should be able to generate several IP connections which can work in two modes: unitary and automatic. The receiver part which receives IP traffic can work in modes: echoer and absorber.

Another option for network transmission degradation is replaying IP network conditions which were first captured in a certain network structure in certain time interval. In this case it is possible to repeat the experiment several times in the same conditions. Several tools are able to capture and replay network conditions (e.g.: *ShunraCloud Catcher* [14], *IP Traffic* [17], etc.)

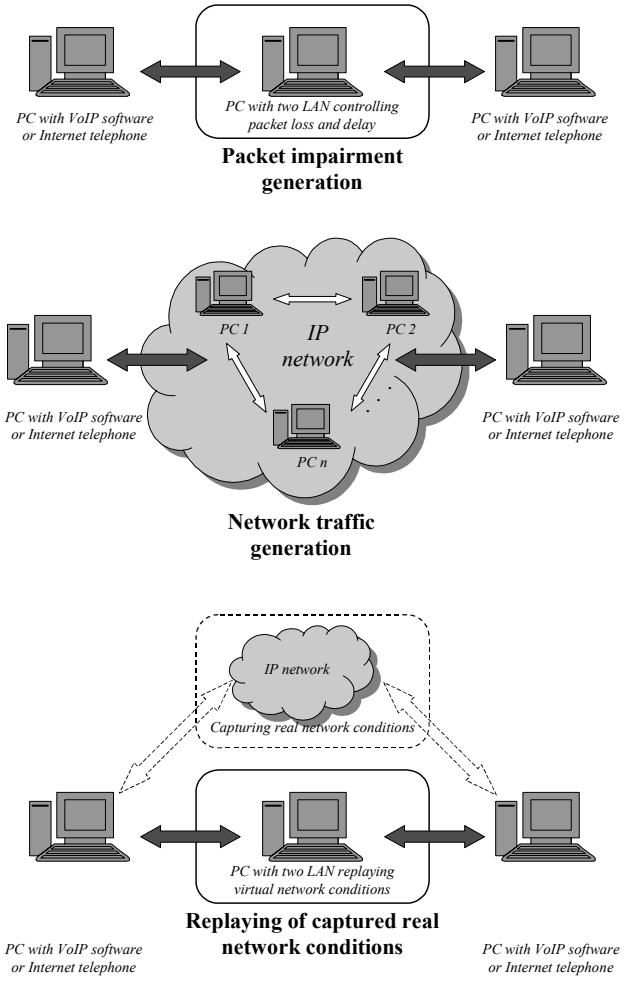



**Figure 2**: Schemes of possible methods of network transmission degradation: direct packet impairment generation, network traffic generation and catching/replaying real network conditions

## 3 SYSTEM SCHEME AND TOOLS

The scheme of voice degradation in VoIP transmission used during tests is presented in Fig. 3. The scheme makes it possible to do voice degradation on the three levels:

1. Acoustics
2. Coding
3. Internet packet transmission (packet loss/delay)

As a reproduction and recording tool a PC equipped with high quality audio interfaces was applied. On the sender part of the scheme the signal is given through the amplifier (Sony TA-F120) to the Bruel&Kjaer telephonometrics system which is responsible for the acoustic degradation of the signal. The Bruel&Kjaer artificial head type 4602 (which consists of the artificial mouth type 4227 and the artificial ear type 4195) fulfils all the technical demands of the ITU-T recommendations P.50 and P.51 [11,12]. The artificial ear has the acoustic impedance which is similar to the human one during a telephone reception and the artificial mouth has the characteristics similar to the human one and emits undistorted sound of the demanded acoustic pressure in the frequency band 100Hz-10kHz. The coder (responsible for coding degradation of the signal) is implemented in the independent telephone software. The applied VoIP telephone set Snom100 [15] can work in two standards: H.323 (ITU-T standard) and SIP (Session Initiation Standard – IETF standard) and uses several codecs (μ-law, a-law, GSM, G.729A). The packet impairment was done on a PC equipped with two LAN interfaces and a packet impairment generator. Apart from packet impairment generation, the *ShunraCloud* [14] lets it also reply network conditions.

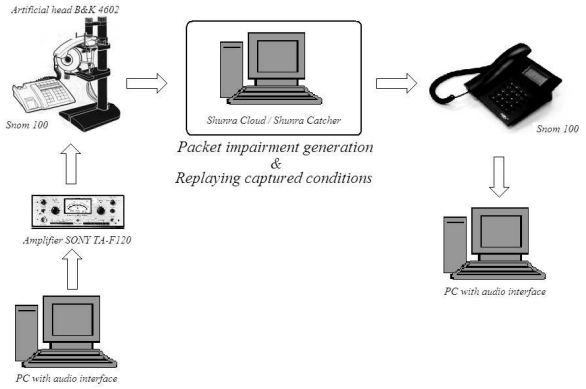



**Figure 3**: Scheme of voice degradation in VoIP transmission applied during tests.

## 4. PRELIMINARY SPEAKER RECOGNITION TESTS

The preliminary tests presented below were carried out for three network conditions modeled with the *ShunraCloud* software. Packet impairment was done with Gilbert-Elliot model (in contrast to *ShunraCloud*, another packer impairment generator, *NetDisturb* offers only the simple Bernoulli model). The Gilbert-Elliot model is a modification of the traditional two states Gilbert model, but it has set BER (Bit Error Rate) for each state: $e_G$ in state "0" and $e_B$ in state "1" (Fig. 2) ($e_G \ll e_B$). Probabilities $p$ and $q$ are defined as in the traditional Gilbert model. For the 1st network conditions: $p$=5.76%, $q$=94% and for the 2nd: $p$=9% and $q$=70%. The proposed conditions (determined by $p$ and $q$) were taken from real network measures of packet transmissions [8] and are close to average or good network conditions. The applied system is able to capture the IP traffic in real time, so the captured traffic can be then replayed. The recorded parameters of IP traffic were done in the following 3rd network conditions:

1. 1st host:156.17.30.34
2. 2nd host: www.shunra.com
3. Time interval: 900 sec.
4. Packet length: 32 bytes
5. Latency (Min/Avg/Max): 60/62/160
6. Average packet loss: 5.26%

For the presented above semi-real IP network conditions system the preliminary tests on a basic speaker recognition application were carried out. Speaker identification on the closed-set of 10 speakers was done with the pole-zero linear prediction features [1] giving better spectral representation than traditional all-poles model. Average orthogonal prediction models not requiring time normalization were estimated for each voice [1]. The results of speaker recognition tests for the three network conditions described above and no degradation (without additional acoustic, coding and Internet package transmission degradations) are presented in Table 3. Beside speaker recognition, the table also presents the results for the subjective assessment of the transmitted signal quality: logatom intelligibility results for the tested conditions. The mediocre results obtained for the degraded signal is probably caused mainly by the acoustic level [5] (the poor quality of the microphone used in the handset). The presented results based on not a very sophisticated speaker recognizer have a preliminary character and include all three degradation levels. Future works will be more focused on separate aspects and decisive factors of speech transmission over IP.

| **1st network conditions** | **2nd network conditions** | **3rd network conditions** | **No degradation** |
|---|---|---|---|
| Speaker recognition correctness | | | |
| 53% | 46% | 64% | 98% |
| Logatom intelligibility | | | |
| 56% | 52% | 60% | 89% |

**Table 3**: The speaker recognition and logatom intelligibility results for the three tested network conditions (G.711, 20ms packet length).

Apart from voice recognition, methods of VoIP transmission presented in the paper, speech degradation can also be used in other applications, like speech recognition or quality assessment over IP. The full simulation of the VoIP transmission channel enables one the full control of the parameters and simulated conditions of the experiment but does not guarantee the full reliability of above. On the other hand, providing semi-real conditions according to the methods proposed above comes across several obstacles, i.e. difficulties with controlling the transmission parameters or a problem of repeating a certain experiment in the same conditions. Additionally, semi-real conditions are usually more expensive and time-demanding, however, they turn out to be more reliable and bring us closer to real problems which can occur during media transmission over a channel.

## 5. REFERENCES


1. Atal B., Schroeder M.R. “Linear prediction analysis of speech based on a pole-zero representation”, J. Acoust. Soc. An. 64(5), Nov.1978, 1310-1318.
2. Basciuk K., Brachmanski S. “The automation of the subjective measurement of logatom intelligibility”, Proc. Audio Eng. Soc., 102nd Conv., AES, Munich, March 1997.
3. Besacier L., Mayorga P., Bonastre J.B., Fredouille C. “Methodology for evaluating speaker robustness over IP networks”, Proc. of a COST 275 workshop The Advent of Biometrics on the Internet, Rome, Italy, November 7-8, 2002, pp. 43-46.
4. Bolot J., Fosse-Parisis S., Towsey D. “Adaptive FEC-Based Error Control for Interactive Audio in the Internet”, Proc. IEEE Infovom’99, 1999.
5. Evans N., Mason J., Auckenthaler R., Stamper R., „Assesment of speaker verification degradation due to packet loss in context of wireless devices“,Proc. of a COST 275 workshop The Advent of Biometrics on the Internet, Rome, Italy, November 7-8, 2002, pp. 43-46.
6. Jiang, W., Schlzrinne, H., „Modeling of Packet Loss and Delay and Their Effect on Real-Time Multimedia Service Quality“, Proc. The 10th International Workshop on Network and Operating System Support for Digital Audio and Video, Chapel Hill, USA, 2000.
7. Staroniewicz P., “Creation of real conditions VoIP database for Speaker Recognition Purposes”, Proc. Second Cost 275 Workshop. Biometrics on the Internet. Fundamentals, Advances and Applications. Vigo, Spain, 2004, pp.23-26.
8. Yanik M., Moon S., Kurose J., Towsey D. “Measurement and modeling of the temporal dependence in packet loss”, Proc. of IEEE Infocom’99, 1999, pp.345-352.
9. ISI, http://www.isi.edu/nsnam, *Network simulator*.
10. ITU-T Recommendation H.323, Packet-based multimedia communications systems.
11. ITU–T Recommendation P.50 (1993), Artificial Voice.
12. ITU–T Recommendation P.51 (1996), Artificial Mouth.
13. “Requirements and methods for logatom intelligibility assessment”, Polish standard PN-90/T-05100.
14. ShunraCloud Wan Emulator, http://www.shunra.com 2002.
15. “Snom 100, Version 1.15 – User’s Manual”, 2002 Snom Technology Aktiengesellschaft.
16. Voice over packet. An assessment of voice performance on packet networks, 2001 Nortel Networks White Paper.
17. ZTI, http://www.zti-telecom.com, *IP Traffic*, Traffic generator and measurement tool for fixed or wireless IP networks.
18. ZTI, http://www.zti-telecom.com, *LanTraffic*, UDP and TCP traffic generation tool.
19. ZTI, http://www.zti-telecom.com, *NetDisturb*, Impairment tool for IP network.

# EFFECT OF IMPOSTOR SPEECH TRANSFORMATION ON AUTOMATIC SPEAKER RECOGNITION

*Driss Matrouf, Jean-François Bonastre, Jean-Pierre Costa*

LIA, Université d'Avignon
Agroparc, BP 1228
84911 Avignon CEDEX 9, France
{driss.matrouf,jean-francois.bonastre,jean-pierre.costa}@lia.univ-avignon.fr

**ABSTRACT**

This paper investigates the effect of voice transformation on automatic speaker recognition systems performance. We focus on increasing the impostor acceptation rate, by modifying the voice of an impostor in order to target a speci c speaker. This paper is based on the following idea: in several forensic situations, it is reasonable to think that some organizations have a knowledge on the speaker recognition method used by the police department and could impersonate a given, well known speaker. This paper presents some preliminary results, showing an increase of about 2.7 time of the likelihood ratio is quite easy to obtain, without a degradation of the natural aspect of the voice.

## 1. INTRODUCTION

Speech is a compelling biometric for several reasons. The two main ones are the natural and easy to produce aspects of speech and the availability of this media in a large spectra of applications. Even if this biometric modality presents lower performance compared to ngerprints or iris modalities, the progress achieved during the 30 last years on intensive work in the area (and particularly during the last decade) bring the automatic speaker recognition systems at a usable level of performance, for commercial applications. During the same period, in the forensic area, judges, lawyers, detectives, and law enforcement agencies have wanted to use forensic voice authentication to investigate a suspect or to con rm a judgment of guilt or innocence [1][2]. Despite the fact that the scienti c basis of person authentication by his/her voice has been questioned by researchers [3][4][5] and the "need of caution" message sent by the scienti c community in [6], forensic speaker recognition methods are largely used, particularly in the context of terrorism world events. Some recent developments show the interest of Bayesian based methods in forensic speaker recognition [7][8][9]. This approach allows to present more precisely the results of a voice identi cation to a judge. If it is a real progress, it does not solve several problems linked on how the method is evaluated, how hypotheses are de ned or how the con dence on the expert is taken into account.
This paper investigates a different point: if you know which identi cation method will be used, if you know the voice of the person *X*, is it possible to transform the voice of someone else, in order to obtain a positive identi cation using the given system and the reference of *X* ? Of course, the transformed voice should correspond to a natural voice.

In this paper we propose a Gaussian-Dependent Filtering technique allowing impostor signal transformation in order to increase the impostor speech likelihood given a targeted speaker. The proposed technique is based on the use of two parallel models (GMM) of the targeted speaker. The rst model corresponds to the one used in the speaker veri cation system and is the master model for frame-based Gaussian component *a posteriori* probabilities computation. The second one is used to estimate the lters to apply to each impostor signal frame. Synthesis of the transformed signal is done frame by frame using the standard overlap-add technique with Hamming windows.

This paper is organized as follow. Firstly, the speaker recognition framework is presented in section 2 and the proposed voice transformation method is presented in section 3. Secondly, a preliminary set of experiences and the corresponding results are presented in section 4 and 5. Some conclusions are proposed in section 6. Finally, section 7 presents the ongoing works.

## 2. STATISTICAL SPEAKER RECOGNITION

Given a segment of speech $Y$ and a speaker S, the speaker veri - cation task is to determine if $Y$ was spoken by S. This task is often stated as basic hypothesis test between two hypotheses:

$H0$: $Y$ is from the hypothesized speaker S,
$H1$: $Y$ is not from the hypothesized speaker S.

The test used to decide between these two hypotheses is a comparison between a likelihood ratio (LR) and a threshold $\theta$. The test is given by:

$$LR(Y, H0, H1) = \frac{p(Y|H0)}{p(Y|H1)} \quad (1)$$

where Y is the observed speech segment, $p(Y|H0)$ is the probability density function for the hypothesis $H0$ evaluated for Y (also referred to as the "likelihood" of the hypothesis $H0$ given the speech segment Y), $p(Y|H1)$ is the likelihood function for $H1$ and $\theta$ is the decision threshold for accepting or rejecting $H0$. If $LR(Y, H0, H1) > \theta$, $H0$ is accepted else $H1$ is accepted.
In order to compute the two likelihoods $p(Y|H0)$ and $p(Y|H1)$, the speech signal is transformed on a set of feature vectors. The aim of this transformation is to obtain a new representation which is more compact, less redundant, and more suitable for statistical

modeling. Most of the speech parameterizations used in speaker veri cation systems relies on a cepstral representation of speech. The output of this stage is typically a sequence of feature vectors representing the test segment $Y = \{y_1, \ldots, y_T\}$. Mathematically, a model denoted $\lambda_{hyp}$ represents $H0$, which characterizes the hypothesized speaker S in the feature space. This model is learned using an extract of speaker S voice. The model $\lambda_{h\bar{y}p}$ represents the alternative hypothesis, $H1$.
The likelihood ratio statistic is then $\frac{p(Y|\lambda_{hyp})}{p(Y|\lambda_{h\bar{y}p})}$. Often, the logarithm of this statistic is used giving the $logLR$:

$$\Lambda(Y) = log(p(Y|\lambda_{hyp})) - log(p(Y|\lambda_{h\bar{y}p})). \quad (2)$$

The major approach to model the alternative hypothesis ($H1$) is to pool speech from several speakers and train a single model. This model is called the world model. The use of Gaussian mixture model (GMM) is the predominate approach used in (text-independent) speaker veri cation systems [10]:

$$p(x|\lambda) = \Sigma_{i=1}^{i=M} w_i N(x|\mu_i, \Sigma_i) \quad (3)$$

where $w_i$, $\mu_i$ and $\Sigma_i$ are weights, means and covariances associated with the Gaussian components in the mixture. The GMM parameters are estimated using the EM algorithm. In this paper we used GMM with diagonal matrix covariances. The speaker model $\lambda_{hyp}$ is generally obtained by adapting the world model. This speaker-adaptation is performed by using some speech data from that speaker and MAP adaptation technique. In the speaker veri cation framework only means of world model are adapted and the other parameters remain unchanged [11].

## 3. SPEECH TRANSFORMATION

Our goal in this paper is to transform speech signal belonging to a speaker S' in order to increase its likelihood given the GMM corresponding to another speaker S. The hypothesis test described before will be disturbed and should be increased. By listening to the resulting signal, the effects of the transformation must appear as natural as possible. Let $Y$ be the signal to be transformed. $Y$ is a set of frames of 30ms following a 10ms frame rate: $Y = \{y1, \ldots, y_n\}$. Each frame $y_i$ can be transformed independently and the resulting signal can be obtained from the transformed frames by using the "overlap-add" technique. It consists in computing the transformed speech frame by frame, using overlapped windows and by adding the resulting window-based signals in order to obtain the nal transformed signal.
The main constraint of this approach is to perform the speech transformation in the spectral domain when the objective is to move the signal in the targeted automatic speaker recognition (ASR) space. In order to achieve this objective, we use two parallel sets of acoustic models for a speaker $S$, the rst one is a cepstral-based GMM (with $\Delta$ and $\Delta^2$ cepstrum), following the state-of-the-art ASR systems (cepstral-based GMM system with cepstral mean subtraction and variance normalization). This ASR model is used to estimate the *a posteriori* probabilities of each of the ASR model Gaussian component given a speech vector frame. The second one, denoted here " ltering" model ( l), is used for computing the optimal time-varying lters parameters. It is also a cepstral-based GMM but it allows to comeback easily in the spectral domain (no mean or variance normalization is applied). In this paper, we work with LPCC-based GMMs instead (LPCC: Linear Predictive Cepstral Coding), for example, of a LPC model because LPCC is more suitable for statistical modeling. By using the a posteriori statistics gathered from the ASR model, there is a one-to-one mapping between the two sets of GMM at the the Gaussian level. Let us consider $y$, a frame of the speaker $S'$ and $x$ its corresponding frame of the speaker $S$ (in fact, $x$ is represented by the mean of one GMM component mean).The source- lter model leads to the following relations in the spectral domain:

$$Y(f) = H_y(f)S_y(f) \quad (4)$$

$$X(f) = H_x(f)S_x(f) \quad (5)$$

where $Y$ and $X$ are the spectral representations of $y$ and $x$. $H_y$ and $H_x$ are the transfer functions corresponding to both signals $x$ and $y$; $S_x$ and $S_y$ are the source signals corresponding to $x$ and $y$. It is important to note that the cepstre is nothing other than a compact representation of the transfer function $H$. So, to bring the cepstre of $y$ as close as possible to the one of $x$, it is enough to replace in equation 4 $H_y$ by $H_x$ :

$$Y'(f) = H_x(f)S_y(f) = \frac{H_x(f)}{H_y(f)}Y(f) \quad (6)$$

We call $H_x$ the target transfer function and $H_y$ the source transfer function. If we decide to not modify the phase of the original signal, the lter to be applied to the signal $y$ becomes:

$$H_{yx}(f) = \frac{|H_x(f)|}{|H_y(f)|} \quad (7)$$

In this paper the transfer functions are estimated as follows:

$$H_x(f) = \frac{G_x}{A_x(f)} \quad (8)$$

$$H_y(f) = \frac{G_y}{A_y(f)} \quad (9)$$

where $A_x(f)$ and $A_y(f)$ are the Fourier transforms of the prediction coef cients of the signals $x$ and $y$. $G_x$ and $G_y$ are the gains of the residual signals $s_x$ and $s_y$ ($S_x$ and $S_y$ are the spectral representation of $s_x$ and $s_y$). The source gain and the prediction coef cients of $y$ are obtained directly from $y$. The source gain and the prediction coef cients of $x$ are obtained from the LPCC coef- cients corresponding to the ltering-model Gaussian component having generated the frame $y$ in speaker $S$ model.
This scheme is generalized by using all the components with different *a posteriori* probabilities. The target transfer function is derived from the linear combination of all the ltering GMM means weighted by their *a posteriori* probabilities. The *a posteriori* probabilities are estimated thanks to the ASR GMM.

$$x_{fil} = \Sigma_{i=1}^{M} p(g_{asr}^i|y)\mu_{fil}^i \quad (10)$$

where $x_{fil}$ is the target representation (at the ltering level) of $H_x$, $p(g_{asr}^i|y)$ the *a posteriori* probability of Gaussian component i given the frame $y$. $\mu_{fil}^i$ is the mean of the Gaussian $g_{fil}^i$ corresponding to the Gaussian $g_{asr}^i$ (with the bijection strategy). The target prediction coef cients are estimated from $x_{fil}$ by using a lpcc-lpc transformation. The Figure 1 presents a block diagram of impostor frame transformation. The aim of this signal tranformation is to increase its likelihood with respect to a xed target speaker.

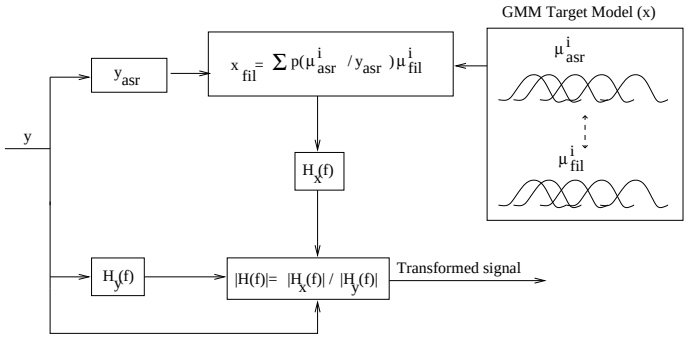



**Fig. 1**. Transform block diagram for one frame: The target transfert function $H_x$ is estimated by using 2 GMM. The srt one allows the *a posteriori* probabilities estimation ; and the second one is used for ltering.

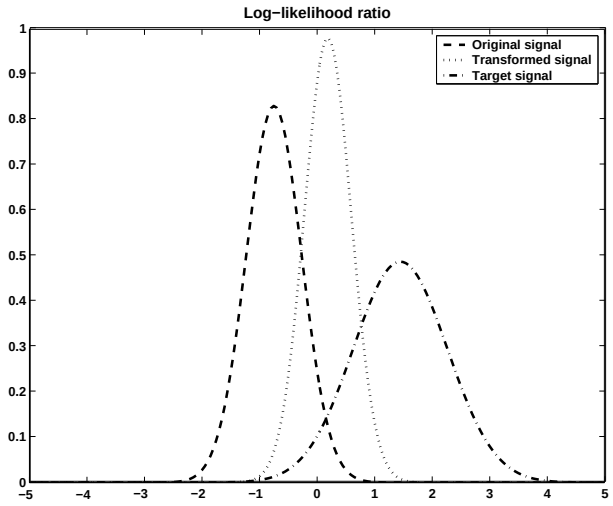



**Fig. 2**. 3 LLR Gaussians. The left one corresponds to LLR of impostor speech segments. The middle one corresponds to LLR of transformed impostor speech segments. The right one corresponds to LLR of target speech segments.

## 4. EXPERIMENTAL PROTOCOL

In this section, we present experiments showing that the described signal transformation can disturb the speaker recognizer system. The training and test data come from ESTER corpus. This acoustic corpus contains shows from 4 different radios recorded between 1998 and 2003. The world GMM are trained by using 100 male speakers which corresponds to 5 hours of speech including silence. 30 male speakers (different form those used for training) are used to be targets (about 1000s). We use PLPCC (Perceptually based Linear Predictive Cepstral Coding) parametrization ; 13 static coef cients (including energy) plus the rst and second derivatives coef cients. On these PLPCC parameters we perform the cepstral mean removal and the normalization of the variance. The PLPCC world GMM contains 1024 Gaussian. The LPCC world GMM is trained in such a way that there is a one-to-one mapping between the two sets of GMM, PLPCC and LPCC: the LPCC world GMM is estimated by using the statistics of the last EM iteration in estimating the PLPCC world GMM. The target speakers PLPCC GMM are estimated by adapting only means of the PLPCC world GMM. Its corresponding LPCC target GMM are estimated by adapting means of the LPCC world GMM but using *a posteriori* probabilities computed in the PLPCC domain.

## 5. RESULTS

In this paper, only the likelihood of speech segments given the target model were optimized. However, the experimental results are based on the log-likelihood ratio (LLR). In the Figure 2, we present 3 LLR Gaussians computed on 1450 segment between 2s and 10s. The left one corresponds to LLR of impostor speech segments. The middle one corresponds to LLR of transformed impostor speech segments. The right one corresponds to LLR of target speech segments. We can see that the confusion area between impostor and target Gaussians is much more important for the transformed signals than for the original impostor segments.

A spectrogram of an original impostor speech segment is shown in Figure 3. Its transformed version is shown in the Figure 4.

The LLR (with respect to a xed target speaker) of the original impostor signal is about $-0.7$, and the LLR of the transformed signal is about $0.3$ ; so LLR increasing of about $1.0$ is observed for that impostor segment. Visually, the spectrum of the transformed signal is not corrupted by the transformation. Listening to the transformed signal, we do not note any distortion and the signal seems natural.

## 6. CONCLUSION

In this paper we investigate the effect of arti cially modi ed impostor speech on a speaker recognition system. It seems reasonable to think that an organization which wants to attribute a speech segment to a given - well known - speaker has a knowledge of the speaker recognition system used by a speci c scienti c police department, as well as a general knowledge on the state-of-the-art in speaker recognition. We demonstrate in this paper that, following this hypothesis, it seems relatively easy to transform the voice of someone in order to target a speci c-speaker voice, in terms of the automatic speaker recognition system.
The preliminary experiment proposed in this paper showed that it is possible to increase an hypothesis test ratio (LR) by a factor around 2.7 with a transformed voice remaining natural. This preliminary result have to be con rmed by a complete experiment, including all the step of state-of-the-art (text independent) speaker recognition systems, like the feature normalization and the score normalization steps (TNorm). The natural aspect of the transformed voice has also to be assessed by a scienti c perception experiment.

## 7. FUTURE WORKS

As presented in the section 6, we plan to perform a complete experiment. We will use the framework of NIST-SRE evaluation campaigns in order to measure precisely the effects of voice transformation in false acceptance rates. We will use for this purpose the system LIA_SpkDet [13] used during the past NIST campaigns[12].

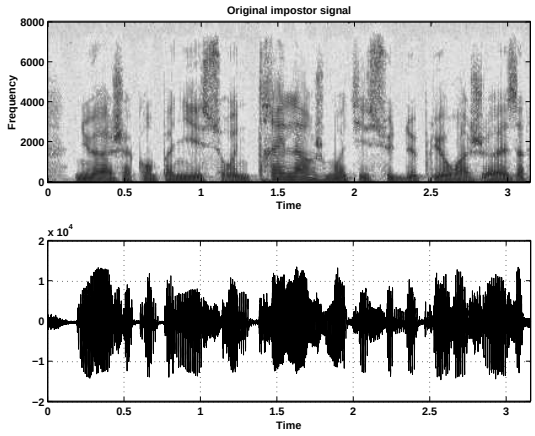



**Fig. 3**. The original impostor signal. The LLR with respect to a xed target speaker is about -0.7.

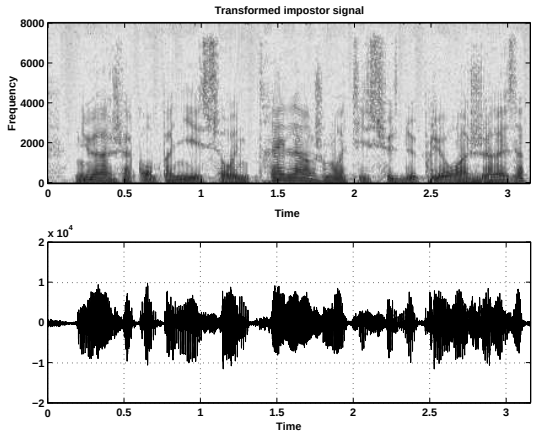



**Fig. 4**. The transformed impostor signal by a Gaussian-Dependent Filtering. The LLR with respect to the same target speaker as in Figure 3 becomes 0.3.

This system is distributed following an open source licence and is based on the ALIZE toolkit, also available with the same licensing system [14][15]. The experiments should explore the effect of the different level of knowledge on the speaker recognition system: knowledge of the method only, knowledge on the feature extraction/normalization process, knowledge on the world model, knowledge on the targeted speaker model...

This experiment should be completed by a perception study, in order to measure when a transformed voice remains natural for a naive auditor and when the voice seems arti cial.

Finally, some improvement on the voice transformation could be explored, by maximizing directly the LLR instead of the client likelihood, or by smoothing the transformation on the entire signal instead to apply it on a frame-based approach. It is also possible to apply a constrained transformation, as the LLR increasing factor obtained in this paper is really large, in order to save the natural aspects of the transformed voice. Finally, it is certainly interesting to extend the approach for taking into account other information, like prosody.

## 8. REFERENCES


[1] R.H. Bolt, F.S. Cooper, D.M. Green, S.L. Hamlet, J.G. McKnight, J.M. Pickett, O. Tosi, B.D. Underwood, D.L. Hogan, "On the Theory and Practice of Voice Identi cation", *National Research Council, National Academy of Sciences, Washington, D.C.*, 1979.

[2] O. Tosi, "Voice Identi cation: Theory and Legal Applications",*University Park Press: Baltimore, Maryland*, 1979.

[3] R.H. Bolt, F.S. Cooper, E.E.Jr. David, P.B. Denes, J.M. Pickett, K.N. Stevens, "Speaker Identi cation by Speech Spectrograms: A Scientists' View of its Reliability for Legal Purposes",*Journal of the Acoustical Society of America*, 47, 2 (2), 597-612, 1970.

[4] J.F. Nolan, "The Phonetic Bases of Speaker Recognition",*Cambridge University Press: Cambridge*, 1983.

[5] L.J. Boë, "Forensic voice identi cation in France",*Speech Communication, Elsevier*, Volume 31, Issues 2-3, June 2000, pp. 205-224 (http://dx.doi.org/10.1016/S0167-6393(99)00079-5).

[6] J.-F. Bonastre, F. Bimbot, L.-J. Boe, J.P. Campbell, D.A. Reynolds, I. Magrin-Chagnolleau, "Person Authentication by Voice: A Need for Caution", *Proceeding of Eurospeech 2003*, 2003

[7] C. Champod, D. Meuwly, "The inference of identity in forensic speaker recognition",*Speech Communication*, Vol. 31, 2-3, pp 193-203, 2000

[8] J. González-Rodr´guez, J. Ortega, and J.J. Lucena, "On the Application of the Bayesian Framework to Real Forensic Conditions with GMM-based Systems", *Proc. Odyssey2001 Speaker Recognition Workshop*, pp. 135-138, Crete (Greece), 2001

[9] P. Rose, T. Osanai, Y. Kinoshita, "Strength of Forensic Speaker Identi cation Evidence - Multispeaker formant and cepstrum based segmental discrimination with a Bayesian Likelihood ratio as threshold", *Speech Language and the Law*, 2003; 10/2: 179-202.

[10] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, D. A. Reynolds, "A tutorial on text-independent speaker veri cation", *EURASIP Journal on Applied Signal Processing*, 2004, Vol.4, pp.430-451

[11] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker veri cation using adapted Gaussian mixture models", *Digital Signal Processing (DSP), a review journal Special issue on NIST 1999 speaker recognition workshop*, vol. 10(1-3), pp 19-41, 2000.

[12] NIST Speaker Recognition Evaluation campaigns webd site, http://www.nist.gov/speech/tests/spk/index.htm

[13] LIA_SpkDet system web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/LIA_RAL

[14] J.-F. Bonastre, f. Wils, S. Meignier, "ALIZE, a free toolkit for speaker recognition", *Proceedings of ICASSP05*, Philadelphia (USA), 2005

[15] ALIZE project web site, http://www.lia.univ-avignon.fr/heberges/ALIZE/

# ON THE USE OF DECOUPLED AND ADAPTED GAUSSIAN MIXTURE MODELS FOR OPEN-SET SPEAKER IDENTIFICATION

*J. Fortuna, A. Malegaonkar, A. Ariyaeeinia and P. Sivakumaran**

University of Hertfordshire, Hatfield, UK, *Canon Research Centre Europe Ltd., Bracknell, UK
{j.m.r.c.fortuna,a.m.ariyaeeinia,a.malegaonkar}@herts.ac.uk, *siva@cre.canon.co.uk

## ABSTRACT

This paper presents a comparative analysis of the performance of decoupled and adapted Gaussian mixture models (GMMs) for open-set, text-independent speaker identification (OSTI-SI). The analysis is based on a set of experiments using an appropriate subset of the NIST-SRE 2003 database and various score normalisation methods. Based on the experimental results, it is concluded that the speaker identification performance is noticeably better with adapted-GMMs than with decoupled-GMMs. This difference in performance, however, appears to be of less significance in the second stage of OSTI-SI where the process involves classifying the test speakers as known or unknown speakers. In particular, when the score normalisation used in this stage is based on the unconstrained cohort approach, the two modelling techniques yield similar performance. The paper includes a detailed description of the experiments and discusses how the OSTI-SI performance is influenced by the characteristics of each of the two modelling techniques and the normalisation approaches adopted.

## 1. INTRODUCTION

Given a set of registered speakers and a sample utterance, open-set speaker identification is defined as a two stage problem [1]. Firstly, it is required to identify the speaker model in the set, which best matches the test utterance. Secondly, it must be determined whether the test utterance has actually been produced by the speaker associated with the best-matched model, or by some unknown speaker outside the registered set. The first stage is responsible for generating open-set identification error (OSIE). The decisions made in the second stage can generate either an open-set identification-false alarm (OSI-FA) or an open-set identification-false rejection (OSI-FR). This paper is concerned with open-set identification in the text-independent mode in which no constraint is imposed on the textual content of the utterances. It is well known that this is the most challenging class of speaker recognition. Open-set, text-independent speaker identification (OSTI-SI) is known to have a wide range of applications in such areas as document indexing and retrieval, surveillance, screening, and authorisation control in telecommunications and in smart environments.

One of the key issues in designing an OSTI-SI system is the selection of the type of speaker modelling technique. The Gaussian mixture model (GMM)-based approach is the most common choice for this purpose. In this technique, a speaker can be modelled by using either a decoupled-GMM [2] or an adapted-GMM [3]. In the former case, each model is built independently by applying the expectation maximisation (EM) algorithm to the training data from a specific speaker. In the latter case, each model is the result of adapting a general model, which represents a large population of speakers, to better represent the characteristics of the specific speaker being modelled. This general model is usually referred to as world model or universal background model (UBM). The common method used for the purpose of adaptation is based on the *maximum a posteriori* (MAP) estimation [4].

Two previous studies by the authors have independently investigated the performance of OSTI-SI using decoupled and adapted models respectively [1][5]. This paper combines the results obtained in the said previous studies and presents a comparative analysis on the use of decoupled and adapted Gaussian mixture models for the purpose of OSTI-SI.

The remainder of the paper is organised in the following manner. The next section describes the speech data, the feature representation and the GMM topologies used in the investigations. Section 3 details the testing procedure adopted, and Section 4 gives a summary of the score normalisations adopted. Section 5 provides a comparative analysis of the results obtained, and the overall conclusions are presented in Section 6.

## 2. EXPERIMENTAL CONDITIONS

The speech data adopted for this comparative study is based on a scheme developed for the purpose of evaluating OSTI-SI [1]. It consists of speech utterances extracted from the 1-speaker detection task of the NIST Speaker Recognition Evaluation 2003. In total, the dataset includes 142 known speakers and 141 unknown speakers. The training data for each known speaker model consists of 2 minutes of speech and each test token from either population contains between 3 and 60 seconds of speech. These amount to a total of 5415 test tokens (2563 for known speakers and 2852 for unknown speakers). Achieving this number of test tokens is based on a data rotation approach which is detailed in [1]. For training the 2048 mixtures of the world model, all the speech material from 100 speakers is used (about 8 hours of speech). In the dataset there are also 505 development utterances from 33 speakers which can be used for score normalisation purposes.

In this study, each speech frame of 20ms duration is subjected to a pre-emphasis and is represented by a $16^{th}$ order linear predictive coding-derived cepstral vector (LPCC) extracted at a rate of 10ms. The first derivative parameters are calculated over a span of seven frames and appended to the static features. The full vector is subsequently subjected to cepstral mean normalisation.

The GMM topologies used to represent each enrolled speaker in the studies involving decoupled and adapted models are 32m and 2048m respectively, where *N*m implies *N* Gaussian mixture densities parameterised with a mean vector and diagonal covariance matrices. In the case of the decoupled models, the parameters of each GMM are estimated using the maximum likelihood (ML) principle through a form of the expectation-maximisation (EM) algorithm [2]. In this case, an initial estimate of the model parameters for the EM algorithm is obtained by using a modified version of the LBG procedure, termed distortion driven cluster splitting (DDCS) [6]. In the case of the adapted models, the parameters of each GMM are estimated from the world model using a form of the MAP estimation procedure [3].

## 3. TESTING PROCEDURE

In each test trial, first, the following are obtained.

$$S_{\mathrm{ML}} = \max_{1 \le n \le N} \left\{ \log\left( p(\mathbf{O} \mid \boldsymbol{\lambda}_n) \right) \right\}, \tag{1}$$

$$n_{\mathrm{ML}} = \arg\max_{1 \le n \le N} \left\{ \log\left( p(\mathbf{O} \mid \boldsymbol{\lambda}_n) \right) \right\}, \tag{2}$$

If $\mathbf{O}$ is originated from the $m^{\mathrm{th}}$ registered speaker and $n_{\mathrm{ML}} \neq m$ then an OSIE is registered and the score discarded. Otherwise, $S_{\mathrm{ML}}$ is normalised (with one of the score normalisation techniques considered in the next section) and stored in one of two groups depending on whether the observation is originated from a known or an unknown speaker. After the completion of all the test trials in a given investigation, the stored $S_{\mathrm{ML}}$ values are retrieved to form the empirical score distributions for both known and unknown speakers. These distributions are then used to determine the open-set identification equal error rate (OSI-EER), i.e. the probability of equal number of OSI-FA and OSI-FR.

When $\boldsymbol{\lambda}_n$ is a decoupled-GMM, the log-likelihood score for the sample utterance $\mathbf{O}$ as shown in (1) is computed as:

$$\log p(\mathbf{O} \mid \boldsymbol{\lambda}_n) = \sum_{t=1}^{T} \left[ \log\left( \sum_{c=1}^{N} w_c^{\lambda_n} b_c^{\lambda_n}(\mathbf{o}_t) \right) \right], \tag{3}$$

where $w_c^{\lambda_n} b_c^{\lambda_n}$ represents the weighted Gaussian probability density function for the $c^{th}$ mixture in the $n^{th}$ speaker model (or world model), $N$ is the total number of mixtures in the speaker models and the world model respectively and $T$ is the number of observations $\mathbf{o}_t$ in each test trial.

When $\boldsymbol{\lambda}_n$ is an adapted-GMM, the score is computed as:

$$\log\left( p(\mathbf{O} \mid \boldsymbol{\lambda}_n) \right) = \sum_{t=1}^{T} \left[ \log\left( \sum_{c=1}^{C} w_{\phi(c,t)}^{\lambda_n} b_{\phi(c,t)}^{\lambda_n}(\mathbf{o}_t) \right) \right], \tag{4}$$

where $w_{\phi(c,t)}^{\lambda_n} b_{\phi(c,t)}^{\lambda_n}$ represents the weighted Gaussian probability density function for the mixture given by $\phi(c,t)$ in the $n^{th}$ speaker model (or in the world model). The function $\phi(c, t)$ represents the indexes of the $C$ mixtures yielding the highest weighted probabilities for the feature vector $\mathbf{o}_t$ in the world model.

## 4. SCORE NORMALISATIONS

The scores computed according to equations (3) and (4) are affected by three main factors: distortions in the characteristics of the test utterance, misalignment of speaker models due to differences in the training conditions, and the problem of unseen data [3]. In order to tackle these problems, score normalisation methods can be used. The normalisations considered in this study are the world model normalisation (WMN), the cohort normalisation (CN), the unconstrained cohort normalisation (UCN), T-norm and various forms of Z-norm. Further details about these methods in the context of OSTI-SI can be found in [1].

## 5. EXPERIMENTAL RESULTS

Table 1 presents the results obtained for the considered modelling techniques in the first stage of the OSTI-SI. These results clearly show that the adapted-GMMs performed significantly better than the decoupled-GMMs. It appears that the coupling between the world model and each adapted-GMM seems to help the first stage of the OSTI-SI because of the better handling of the unseen data [3] as well as the contaminations of the test data.

| | **Decoupled-GMMs** | **Adapted-GMMs** |
|---|---|---|
| OSIE (%) | 33.7 ± 1.8 | 27.0 ± 1.7 |

**Table 1:** Relative performance of the considered modelling techniques in the first stage of OSTI-SI. The error rates are given with a 95 % confidence interval.

Table 2 shows the performance of the considered modelling techniques in the second stage of the OSTI-SI with various score normalisation methods. It also includes relative effectiveness of these modelling techniques without any form of score normalisation i.e. when the likelihood scores were determined according to equations (3) and (4).

| **Normalisation** | **Decoupled-GMMs** | **Adapted-GMMs** |
|---|---|---|
| None | 43.6 ± 2.4 | 47.8 ± 2.3 |
| WMN | 29.6 ± 2.2 | 22.9 ± 1.9 |
| $\mathrm{WMN_Z}$ | 26.8 ± 2.1 | 20.7 ± 1.8 |
| CN | 22.5 ± 2.0 | 20.7 ± 1.8 |
| $\mathrm{CN_Z}$ | 20.9 ± 1.9 | 19.1 ± 1.8 |
| UCN | 19.1 ± 1.9 | 18.5 ± 1.8 |
| $\mathrm{UCN_Z}$ | 20.7 ± 1.9 | 18.3 ± 1.8 |
| T-norm | 34.2 ± 2.3 | 18.6 ± 1.8 |
| TZ-norm | 29.6 ± 2.2 | 18.0 ± 1.7 |

**Table 2:** Results obtained in the second stage of the OSTI-SI (results are given in terms of OSI-EER(%) with a 95% confidence interval).

These results indicate that without any form of normalisation the use of adapted-GMMs leads to a higher error rate than that obtained with the decoupled-GMMs. This is thought to be due to the effect of the speaker independent components in each adapted-GMM. It should be noted that such an effect can be removed by using WMN and therefore it is common in the literature to consider the performance of the adapted-GMMs in conjunction with WMN as the baseline [3].

Table 2 shows that the adoption of WMN results in a significantly better result for the adapted-GMMs than for the decoupled-GMMs. Figure 1, which shows the DET curves obtained for WMN with these two modelling techniques, further confirms this relative effectiveness. At the first glance, it may

be thought that this difference in performance is solely due to the better handling of the unseen data in the case of adapted-GMMs. However, it can be argued that in the second stage of OSTI-SI, this problem exists to a lesser extent. This is because a speaker model selected in the first stage is always the best match for the test utterance over all the registered speaker models. It is therefore felt that the difference in the observed performance is too significant for it to be solely attributed to the better handling of the unseen data by the adapted-GMM. It is thought that different GMM topologies for the speaker models and the world model could contribute to this difference.

It can be realised from Section 2 that, in the case of decoupled-GMMs, such a topological difference does exist. In this case, the speaker models are built with 32 mixtures whilst the world model consisted of 2048 mixtures. It is believed that with such a degree of topological difference, the contaminations in the test utterance could be reflected very differently in the best matched speaker model and the world model, compared to that in the case where the relevant models are of unique topology (which has been the case in adapted-GMMs). As a result, in the case of decoupled-GMMs, WMN may not be as effective as it is in the case of adapted-GMMs in compensating for such contaminations in the test utterance. In order to verify this hypothesis, a world model with 32 mixtures was trained using the same speech data as that for the 2048 mixture version. Table 3 presents the result of this study. It can be seen that, in this case, the performance of WMN improves significantly.

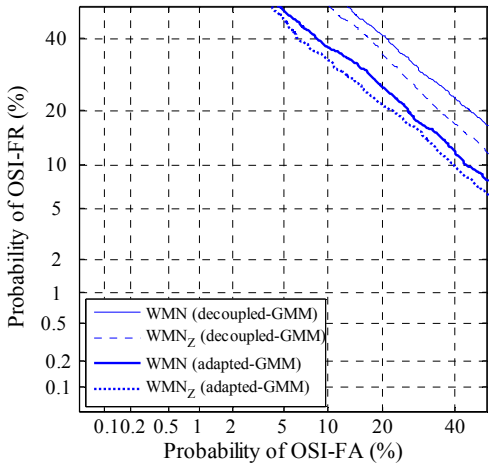



**Figure 1:** DET plots for the considered modelling techniques with WMN and WMN$_Z$

| | World Model Topology | |
|---|---|---|
| | **2048m** | **32m** |
| **OSI-EER (%)** | 29.6 ± 2.2 | 24.2 ± 2.0 |

**Table 3:** Effectiveness of the WMN for two different world model topologies (in the case of decoupled-GMM). (Results are given with a 95 % confidence interval).

Table 2, Figure 2 and Figure 3 indicate that in the cases of CN and UCN, the decoupled-GMMs offers similar levels of performance to those obtainable with the adapted-GMMs. It is also observed that the performance of the decoupled-GMMs followed that of the adapted-GMMs more closely in the case of UCN than in the case of CN. When the adapted-GMMs are used with CN/ UCN, the cohort speaker models have to take the role of handling the unseen data. These models cannot be as effective as the world model in accomplishing this task. This is because, in the case of CN and more in the case of UCN, there is no guarantee that the unseen data falls outside the adapted regions of the competing models. For the same reason, the performance obtained with CN and UCN in adapted-GMMs may not be considerably different from that in decoupled-GMMs. Based on the results obtained for CN and UCN, it appears that the cohort speaker models that are chosen based on their closeness to the best matched speaker model are better in accomplishing this task than the cohort speaker models chosen according to their closeness to the test utterance.

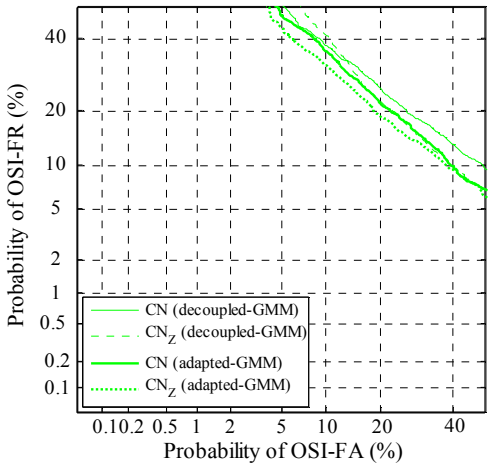



**Figure 2:** DET plots for the considered modelling techniques with CN and CN$_Z$.

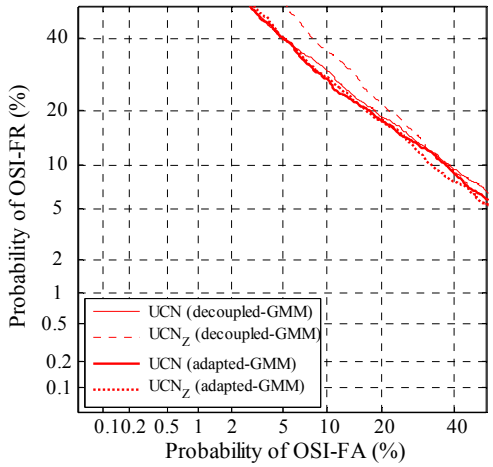



**Figure 3:** DET plots for the considered modelling techniques with UCN and UCN$_Z$.

It is interesting to note in Table 2 that the T-norm approach, which is one the worst performers in the case of the decoupled-GMMs, is one of the best performers in the case of adapted-GMMs. Figure 4 elaborates these results using DET curves. A careful investigation into these results shows that

the reason for this is, to a large extent, dependent on how each registered speaker model reacts to a given test utterance produced by an unknown speaker. In the case of adapted-GMMs, this reaction is much similar across the registered speaker population, whereas in the case of decoupled-GMMs, it is considerably different. As a result, the T-norm parameters computed for adapted-GMMs tend to be much closer to those of the unknown speaker distribution and this makes the T-norm work better in this case. It should be noted that the Z-norm, which is specifically designed for aligning the models (i.e. reducing the model dependant biases), tend to produce more consistent reactions across the registered speaker population to a given test utterance produced by an unknown speaker. This may explain why, in the case of decoupled-GMMs, when T-norm is combined with Z-norm, a relatively large improvement is observed (Table 1 or Figure 4).

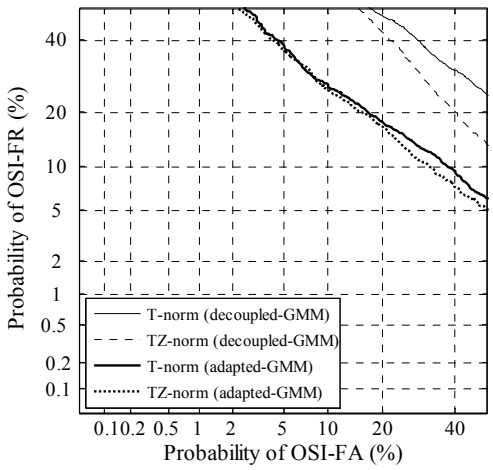



**Figure 4:** DET plots for the considered modelling techniques with T-norm and TZ-norm.

It is observed in Table 2 and Figures 1– 4 that with two exceptions, Z-norm exhibits similar levels of performance for both considered modelling techniques when it is combined with other normalisation methods. These exceptional cases are the T-norm and Z-norm combination (i.e. TZ-norm) which is discussed above, and the UCN and Z-norm combination (i.e. $UCN_Z$).

In the case of decoupled-GMMs, $UCN_Z$ performs slightly worse than UCN. A close analysis of this case revealed that the underlying problem was the lack of availability of sufficient data for computing the Z-norm parameters for every known speaker model. In particular, it was observed that, with the available development data, the tail ends of the distributions assumed for computing the Z-norm parameters were significantly inaccurate. This problem may be tackled by adopting a large development set representing enough varieties of unknown speaker utterances. In other words, for each registered model, there should be an adequately large subset of the development data that can effectively be used as the unknown speaker utterances. Achieving this in practice is extremely difficult, especially when dealing with a large set of registered models. Therefore, it may be best to avoid the use of combined Z-norm and UCN with decoupled-GMMs. However, this problem is not as significant when decoupled-GMMs are replaced with adapted-GMMs. This is because, with adapted-GMMs, the scores produced by registered speakers for unknown utterances (in the development set) tend to be very similar. As a result, for each registered model, the validity of the Z-norm parameters obtained using the relevant subset of the development data is not too significantly influenced by the size of the subset. This may be the reason that, in the case of adapted-GMMs, $UCN_Z$ does not achieve a worse error rate than UCN.

# 6. CONCLUSIONS

This paper has presented a comparative analysis of the performance of decoupled-GMM and adapted-GMM in OSTI-SI. It has been shown that, in general, the use of adapted-GMM results in better performance and this is particularly significant in the first stage of the OSTI-SI process. The better performance of the adapted-GMMs has been mainly attributed to the way in which such models handle the problem of the unseen data in the test segments. It was also found out that significant differences in the model topology limit the effectiveness of the WMN for the case of decoupled models. Furthermore, based on the experimental results it is shown that the cohort approaches are equally capable of achieving good performance with both types of models and this is found to be particularly evident for the case of UCN. It is also noted that T-norm is one of the worst performers in the case of decoupled-GMM despite being amongst the best performers in the case of adapted-GMM. Finally, the performance improvement achievable by Z-norm is similar with both modelling approaches with the exception of the cases involving UCN and T norm (i.e. $UCN_Z$ and TZ-norm).

# 7. REFERENCES


[1] Fortuna, J., Sivakumaran, P., Ariyaeeinia, A. M., and Malegaonkar, A., "Relative Effectiveness of Score Normalisation Methods in Open-set Speaker Identification", Proc. Odyssey 2004 Speaker and Language Recognition Workshop, pp. 369-376, 2004.

[2] Reynolds, D., Rose, R. C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech Audio Proc., vol 3, 1995.

[3] Reynolds, D., Quatieri, T. F., and Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, pp. 19-41, 2000.

[4] Gauvain, J. L. and Lee, C.-H., "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. Speech Audio Process., vol. 2, pp. 291-298, 1994.

[5] Fortuna, J., Sivakumaran, P., Ariyaeeinia, A. M., and Malegaonkar, A., "Open-set Speaker Identification Using Adapted Gaussian Mixture Models", to appear in Proc. Interspeech'05, 2005.

[6] Ariyaeeinia, A. M. and Sivakumaran, P. "Comparison of VQ and DTW classifiers for speaker verification," Proceedings of the IEE European Convention on Security and Detection (ECOS'97), No. 437, pp. 142-146, April 1997.

# HARMONIC DECOMPOSITION FOR ROBUST SPEAKER RECOGNITION

*Bostjan Vesnicer, France Mihelic and Nikola Pavesić*

Faculty of Electrical Engineering, University of Ljubljana, Slovenia
<bostjan.vesnicer@fe.uni-lj.si>

## ABSTRACT

This paper tries to assess the effectiveness of the harmonic decomposition of the speech signal in a speaker veri cation task. The decomposition is achieved by applying the PSHF technique. The performance of the method is evaluated on the XM2VTS speech database. The results show signi cant improvement over the baseline approach if conditions between enrollment data and test data are mismatched.

## 1. INTRODUCTION

One of the main remaining problems in speaker recognition is the possible mismatch between enrollment and test data. Two main factors may contribute to this mismatch: variation in speaker voice due to emotion, health state, or age and environment condition changes in transmission channel, recording material, or acoustical environment [1].

Several normalization techniques have been introduced explicitly to cope with mismatched conditions. Normalization can be performed either in score domain or in feature domain [1]. Various kinds of score normalization techniques have been proposed in the literature, mainly derived from work of Li and Porter [13]. On the other hand, cepstral mean subtraction (CMS), feature variance normalization and feature warping try to perform normalization in feature space.

In contrast to CMS, which tries to remove the contribution of slowly varying convolutive noises, we propose another form of feature level normalization technique based on harmonic decomposition of the speech signal. We believe that harmonic decomposition can improve speaker recognition/veri cation performance in presence of additive noise, as showed previously for speech recognition [6].

The rest of the paper is organized as follows. In the next section the PSHF technique is brie y depicted. In Section 3 the experiments are described. The results are given and commented in Section 4. Some nal remarks and conclusions are given in Section 5.

## 2. HARMONIC DECOMPOSITION

The decomposition of the speech signal into simultaneous periodic and aperiodic components has been successfully applied for different applications [4]. There have been few alternative methods proposed for this task [4]. In our experiments we adopted *pitch-scaled harmonic lter* (PSHF) [4]; on Internet available implementation [5]. The method was originally developed for acoustic analysis of speech [11] and recently tried on a speech recognition task [6].

In sequel, the PSHF algorithm is brie y described, while the interested reader should refer to [4] for the details.

In rst step the PSHF tries to optimize the initial estimate of the pitch frequency of the analysis frame according to the cost function, which is based on the fact that spectrum computed by discrete Fourier transform (DFT) has well-formed sharp harmonics only if the number of points involved in DFT calculation is exactly an integer multiple of the fundamental frequency.

When the fundamental frequency is determined, the input speech signal is windowed with a pitch-scaled Hann window. Then DFT is applied to a windowed signal in order to obtain its spectrum. Only those bins of the spectrum are admitted by the harmonic lter which correspond to the harmonics of the speech signal, while setting the rest of the bins to zero. The resulting spectrum is transformed into time domain by an inverse DFT and rewindowed to compensate for all the frames that contribute to each sample since the windows may overlap. Thus, a periodic estimation is obtained.

Finally, the aperiodic component is obtained simply by subtracting the periodic component from the original waveform.

For illustration purposes, the result of harmonic decomposition of a short utterance is shown on Figure 1.

## 3. EXPERIMENTS

### 3.1. Dataset

The experiments were performed on audio part of the XM2VTS database [8]. The training set, evaluation set and test set were chosen according to Con guration 1 of the Lausanne protocol [8].

### 3.2. Preprocessing

In order to prepare acoustic features for our experiments, the following steps were undertaken.

#### *3.2.1. Pitch detection*

Since the PSHF expects initial estimates of the fundamental frequency, the pitch detection has to be performed prior to harmonic decomposition.

In our experiments we harnessed the Entropic's get_f0 pitch tracker, which is based on autocorrelation and linear prediction and tries to reduce unavoidable errors by heuristics and dynamic programming [10]. The frame rate for pitch extraction was set to 10 ms.

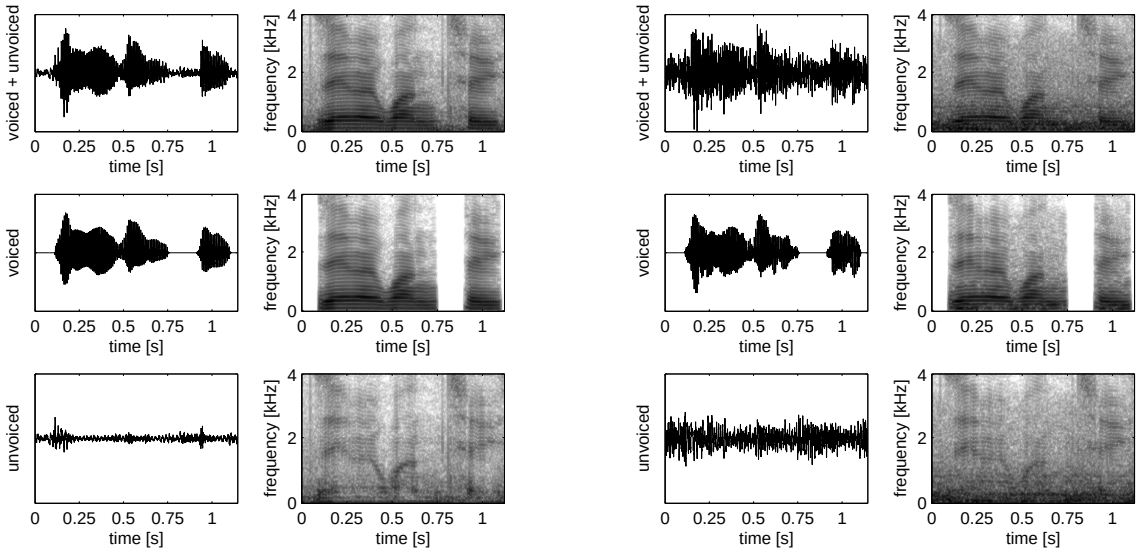



**Figure 1**: Harmonic decomposition of the utterance "zero one two" for clean (left) and noisy (SNR 5dB) speech.

### *3.2.2. Harmonic decomposition*

The above described PSHF algorithm was used for decomposition of the speech into periodic and aperiodic components. The PSHF expects at the input a waveform and its corresponding pitch contour and produces synchronized periodic and aperiodic waveforms at the output. The default settings (4 periods long window, rst 8 harmonics considered by the cost function) were used, except that the internal pitch sampling period was set to 1 ms. The choice for the settings was based on the compromise between performance, checked by informal listening tests, and computational cost.

### *3.2.3. Feature extraction*

Standard 13 dimensional MFCC feature vectors (with logarithm of raw energy instead of the 0th cepstral coef cient) were extracted from original waveforms and periodic waveforms, while aperiodic waveforms were left untouched since we have not used them in our experiments.

The HTK tool HCopy [3] was used for feature extraction. Feature vectors were extracted from 25 ms long Hamming-windowed speech frames at 10 ms frame rate. No cepstral mean normalization was performed since the aim of the experiments was to test the performance of harmonic decomposition solely.

Since the PSHF algorithm operates only on voiced segments of speech, the extracted feature vectors at unvoiced segments contain zeros. These vectors were removed from feature les since they can cause numerical problems at later steps (i.e. training and recognition). To achieve unbiased comparison, all "unvoiced" vectors were removed also from feature les extracted from original waveforms. This is somewhat different from the approach described in [6], where a dither was added to the unvoiced segments in the voiced feature les. As a consequence, no silence detector was needed, since silence frames were effectively removed with "unvoiced" frames.

## 3.3. Training

Gaussian mixture models (GMMs) were used for statistical modeling. First, a universal background model (UBM) was trained using the material from all the speakers from the training set (clients). The training started with estimating mean and variance of a single Gaussian. At each step the number of Gaussians was doubled according to Linde-Buso-Gray (LBG) algorithm and re-estimated with 20 iterations of the maximum likelihood (ML) algorithm. The training procedure stopped when the number of Gaussians reached the value of 256.

The training of the UBM was followed by the adaptation step, where individual speaker models were derived from UBM by maximum a posteriori (MAP) adaptation [7].

## 3.4. Recognition

At the recognition step, every feature vector le from the test set was tested against all client models and in each case the score was obtained as a log likelihood ratio (LLR) between log likelihood of the UBM model and log likelihood of the individual client model. If the score exceeded some threshold, the test utterance was said to belong to the target speaker, while the test utterance was said not to belong to the target speaker, if the score was below this threshold.

The threshold was found by choosing a compromise between false acceptance (FAR) and false rejection (FRR) ratios. If the testing set was used for tuning the threshold, this threshold is said to be a posteriori, while if we use separate evaluation data, the threshold is said to be a priori. Although the a posteriori threshold is unavoidably used when plotting Receiver Operating Characteristic (ROC) or Detection Error Tradeoff (DET) curves [14], it should be noted that a priori threshold gives more realistic expected performance of the system since a testing set is not available in practice [9].

Although it is common to apply one or more score normalization techniques prior to choosing the threshold, we did not use score normalization in our experiments.

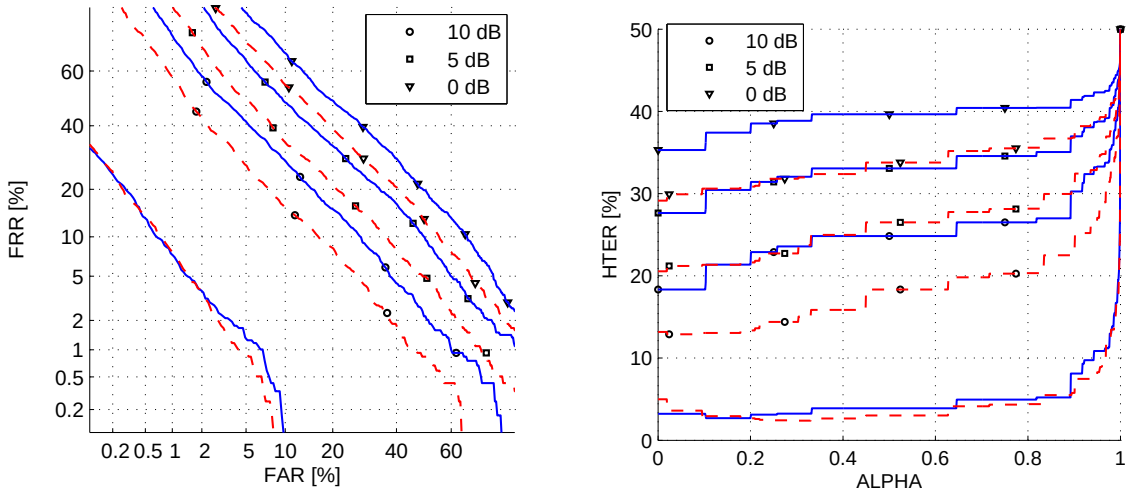



**Figure 2**: Comparison of results for the baseline (solid line) and HDP (dashed line) for noisy (bab) conditions. The lowest two curves in each plot correspond to clean conditions.

## 4. RESULTS

In order to verify our hypothesis that harmonic decomposition can improve performance of speaker veri cation systems in terms of FAR and FRR in presence of additive noise, we compared a system where harmonic decomposition was not performed (baseline), with a system where periodic component of harmonically decomposed signals were used (HDP).

First, the systems were compared in clean conditions with no noise added. After that, two noise signals, namely babbel noise (bab) and automobile highway noise (hwy) [2] were added to all the test utterances using the FaNT tool [12]. We used three different levels of SNR (10 dB, 5 dB and 0 dB). After the addition of each different noise all preprocessing steps needed to be repeated prior to veri cation, but only for the test set.

The results of the experiments are shown in Table 1, Table 2 and Table 3 according to the directions found in Lausanne protocol [8]. Furthermore, the results are also graphically presented in DET and EPC plots (Figure 2, Figure 3). The reader should refer to [9] for the explanation of EPC plots.

| **clean** | FAR | FRR | HTER | |
|---|---|---|---|---|
| $t_{FAE=0}$ | 0.0 | 96.7 | 48.4 | baseline |
| | 0.0 | 95.0 | 47.5 | HDP |
| $t_{FRE=0}$ | 5.1 | 1.3 | 3.2 | baseline |
| | 9.9 | 0.1 | 5.0 | HDP |
| $t_{FAE=FRE}$ | 1.0 | 7.2 | 4.1 | baseline |
| | 1.8 | 4.0 | 2.9 | HDP |
| $\infty$ dB | | | | |

**Table 1**: Comparison of results for the baseline and HDP for clean conditions.

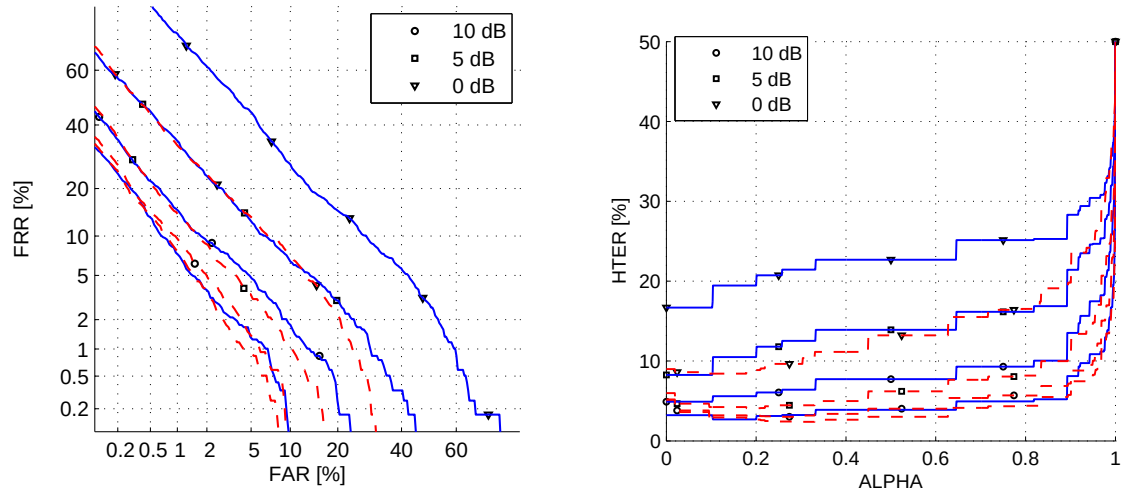



**Figure 3**: Comparison of results for the baseline (solid line) and HDP (dashed line) for noisy (hwy) conditions. The lowest two curves in each plot correspond to clean conditions.

| **bab** | FAR | FRR | HTER | FAR | FRR | HTER | FAR | FRR | HTER | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t_{FAE=0}$ | 0.0 | 98.3 | 49.2 | 0.0 | 97.9 | 49.0 | 0.2 | 96.6 | 48.4 | baseline |
| | 0.0 | 97.4 | 48.7 | 0.0 | 97.0 | 48.5 | 0.1 | 96.6 | 48.4 | HDP |
| $t_{FRE=0}$ | 11.7 | 25.0 | 18.3 | 13.8 | 41.5 | 27.6 | 15.5 | 55.1 | 35.3 | baseline |
| | 15.9 | 10.4 | 13.2 | 18.9 | 22.2 | 20.6 | 22.0 | 36.3 | 29.1 | HDP |
| $t_{FAE=FRE}$ | 3.7 | 47.1 | 25.4 | 5.6 | 61.2 | 33.4 | 7.7 | 71.8 | 39.8 | baseline |
| | 4.2 | 29.7 | 17.0 | 6.3 | 44.9 | 25.6 | 9.3 | 56.8 | 33.1 | HDP |
| | | 10 dB | | | 5 dB | | | 0 dB | | |

**Table 2**: Comparison of results for the baseline and HDP for noisy (bab) conditions.

| **hwy** | FAR | FRR | HTER | FAR | FRR | HTER | FAR | FRR | HTER | |
|---|---|---|---|---|---|---|---|---|---|---|
| $t_{FAE=0}$ | 0.0 | 99.3 | 49.7 | 0.0 | 99.8 | 49.9 | 0.0 | 99.7 | 49.9 | baseline |
| | 0.0 | 98.1 | 49.1 | 0.0 | 99.3 | 49.7 | 0.0 | 100.0 | 50.0 | HDP |
| $t_{FRE=0}$ | 5.2 | 4.6 | 4.9 | 6.8 | 9.8 | 8.3 | 13.9 | 19.4 | 16.7 | baseline |
| | 10.3 | 0.1 | 5.2 | 11.2 | 0.8 | 6.0 | 13.4 | 4.5 | 9.0 | HDP |
| $t_{FAE=FRE}$ | 1.0 | 15.2 | 8.1 | 1.5 | 27.3 | 14.4 | 5.6 | 41.2 | 23.4 | baseline |
| | 1.7 | 5.7 | 3.7 | 1.8 | 9.3 | 5.6 | 2.2 | 22.6 | 12.4 | HDP |
| | | 10 dB | | | 5 dB | | | 0 dB | | |

**Table 3**: Comparison of results for the baseline and HDP for noisy (hwy) conditions.

From Tables 2 and 3, as well as from Figures 2 and 3 it can be seen that:

1. the HDP outperforms the baseline in noisy conditions,
2. performance of both systems drops rapidly by increasing the SNR and
3. the difference between the performance of both systems is dependent on the type of noise.

## 5. CONCLUSIONS

We have proposed a feature-level normalization technique based on harmonic decomposition of the speech signal. The initial results show that the method gives a signi cant performance boost over the baseline approach in presence of additive noise.

Currently, only periodic component of the speech signal was taken into consideration while the information found in aperiodic part was discarded. Since we know that fricatives can be very discriminative along different speakers, it should be useful to nd an appropriate way for combining information from both components.

Furthermore, different nature of both components indicates that we should try to apply distinct parametrization methods for individual constituents.

## 6. REFERENCES


1. F. Bimbot et al., "A Tutorial on Text-Independent Speaker Veri cation", *EURASIP JourApplied Sig. Proc.*, vol. 4, pp. 430–451, 2004.
2. J. H. L. Hansen and L. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit-card corpus", *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 169–184, 1995.
3. The HTKBook, Cambridge University Engineering Department, 2005, http://htk.eng.cam.ac.uk/.
4. P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech", *IEEE Trans. Speech and Audio Processing*, vol. 9,no. 7, pp. 713–726, 2001.
5. P. J. B. Jackson, D. M. Moreno, J. Hernando, and M. J. Russel, "Columbo Project", Uni. of Surrey, http://www.-ee.surrey.ac.uk/Personal/P.Jackson/Columbo/.
6. D. M. Moreno, "Harmonic decomposition applied to automatic speech recognition", M. Sc. thesis, Universitat Politècnica de Catalunya, Barcelona, 2002.
7. D. A. Reynolds, T. F.Quatieri, and R. B. Dunn, "Speaker Veri cation Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41.
8. K. Messer et al., "XM2VTSbd: The Extended M2VTS Database", *Proc. 2nd Conference on Audio and Video-base Biometric Personal Veri cation* , Springer Verlag, New York, 1999.
9. S. Bengio, J. Mariéthoz, and M. Keller, "The Expected Performance Curve", in *International Conference on Machine Learning*, WS ROC Anal. Mach. Learn., 2005.
10. D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", W. .B. Kleijn, K. K. Paliwal, *Speech Coding and Synthesis*, Elsevier Science, str. 495–518, 1995.
11. P. J. B. Jackson, "Characterisation of plosive, fricative and aspiration components in speech production", PhD. Thesis, University of Southampton, UK, 2000.
12. G. Hirsch, "FaNT—Filtering and Noise Adding Tool", Niederrhein University of Applied Sciences, http://dnt.-kr.hsnr.de/download.html.
13. K. P. Li and J. E. Porter, "Normalization and selection of speech segments for speaker recognition scoring", *Proc. ICASSP*, vol. 1, pp. 595–598, New York, 1988.
14. A. Martin et al., "THE DET curve in assessment of detection task performance", *Proc. Eurospeech*, pp. 1895–1898, 1997.

# SPEAKER VERIFICATION USING FAST ADAPTIVE TNORM BASED ON KULLBACK-LEIBLER DIVERGENCE

*Daniel Ramos-Castro, Daniel Garcia-Romero, Ignacio Lopez-Moreno and Joaquin Gonzalez-Rodriguez*

ATVS (Speech and Signal Processing Group)
Universidad Autonoma de Madrid (Spain)
daniel.ramos@uam.es

## ABSTRACT

This paper presents a novel speaker-adaptive score normalization scheme for speaker veri cation based on the widely known Test-Normalization technique, also known as TNorm. This popular method makes use of a cohort of impostor speakers for normalization. The presented technique selects a subset of speakers from the cohort using a distance-based criterion which optimize the adaptation between the target speaker and the normalizing cohort. We have called this technique KL-TNorm, because the distance measure used is a fast approximation of the Kullback-Leibler (KL) divergence for Gaussian Mixture Models (GMM). The low resource cost of this distance computation algorithm allows the use of big cohorts in real applications. Experiments in NIST SRE 2005 show signi cant improvement for the GMM and SVM ATVS-UAM speaker recognition systems for two different training conditions.

## 1. INTRODUCTION

Score normalization techniques have been widely used in recent years to improve speaker veri cation systems performance. These techniques transform the output scores of a system in order to reduce misalignments in the score distributions among different speaker models due to speaker-dependent or speaker-independent factors (such as the conditions of the training or test data, or the mismatch between them). Test-normalization technique, also known as TNorm [1] has been proposed as a test-dependent normalization approach which estimates the score distribution of a speech segment tested using a set of impostor models. In this way, this technique reduces the effect of the test conditions in the score distribution of the speaker models. TNorm has become widely used in the last years mainly due to its good performance at low False Acceptance (FA) rates. Assuming that we have a sequence of feature vectors $O$ extracted from a test utterance and a speaker model, $\lambda_t$. We compute a score $s(O, \lambda_t)$ by testing the observation $O$ with the model $\lambda_t$. TNorm uses a set of impostor models $\Lambda_I = \{\lambda_{I,1}, \cdots, \lambda_{I,N}\}$ to obtain an estimate of the impostor scores $S_I = \{s(O, \lambda_{I,1}), \cdots, s(O, \lambda_{I,N})\}$ associated with the test utterance. The set $\Lambda_I$ is also known as a *cohort* of models. A parametric, ML estimation of the impostor score distribution is performed assuming Gaussian distribution:

$$f(x|S_I) = N(\mu_{TNorm}, \sigma_{TNorm}) \tag{1}$$

and TNorm is then computed in the following way:

This work has been supported by the Spanish Ministry for Science and Technology under project TIC2003-09068-C02-01.

$$s_{TNorm}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{TNorm}}{\sigma_{TNorm}} \tag{2}$$

In this way, TNorm models and normalizes the score distribution of test utterances tested as impostor trials. In this sense, this technique is very similar to cohort-normalization techniques [2], which have been discussed in the past as an alternative to the widely adopted Universal Background Models (UBM) [3]. In cohort normalization, the likelihood computed of a test segment for a given speaker model is normalized by a set of likelihoods computed from impostor models. It has been shown in the literature [2] that these techniques perform better when the cohort is selected in a speaker-dependent way. Speci cally , the more *similar* the cohort models are to the speaker model the better the normalization performance. The way of measuring this similarity between models can be based on qualitative (e. g., transmission channel, sex, environmental conditions) or quantitative criteria. In [2], a distance measure based on likelihood computation is proposed as a quantitative value for model similarity in cohort normalization. Recently, a data-driven distance measure has been proposed for cohort selection in TNorm [4]. In this paper we propose the use of a quantitative distance measure for cohort-selection in TNorm based on a fast approximation of the Kullback-Leibler divergence for Gaussian Mixture Models (GMM) [5]. We have called this normalization technique KL-TNorm. The main advantage of the described distance measure is based on its low resource requirements and its extremely fast performance compared to other methods.

The paper is organized as follows: section 2 describes the fast approximation to Kullback-Leibler divergence for Hidden Markov Models (HMM) presented in [5], and shows its particularization to GMM. In section 3, KL-TNorm is described as an speaker-adaptive TNorm technique. Experiments using the ATVS GMM and SVM spectral systems −presented in NIST SRE 2005 [6]− are described in section 4. The obtained results show that this technique improves the system performance when it is applied to different speaker recognition systems in blind NIST SRE conditions. Finally, in section 5 conclusions are drawn.

## 2. FAST APPROXIMATION TO KULLBACK-LEIBLER DIVERGENCE FOR GMM

The Kullback-Leibler divergence between two probability density functions [8], also known as differential entropy, is de ned by the following expression:

$$D\left(f\left|\hat{f}\right.\right) \equiv \int f \log \frac{f}{\hat{f}} \tag{3}$$

where $f$ and $\hat{f}$ are arbitrary probability density functions (pdfs). This expression can be seen as an asymmetric similarity measure between two random variables. In cases in which it is not possible to compute this expression analytically, it is usually estimated via computer-intensive algorithms such as Monte-Carlo estimation, which usually demand high computational costs, especially when high-dimension distributions are handled. This represents a problem in real applications, where resources requirements are usually critical.

It has been recently shown in the literature [5] that Equation 3 can be upper-bounded by a single expression when two HMM are involved, and therefore we can particularize this expression to the GMM case. Moreover, when two GMM adapted via mean-only-MAP from the same UBM [3] are involved, then such upper-bound can be computed very ef ciently . Let $f = f(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i f_i$ and $\hat{f} = f\left(\mathbf{x}|\hat{\lambda}\right) = \sum_{i=1}^{M} \hat{w}_i \hat{f}_i$ be two pdfs associated with their corresponding GMM models $\lambda$ and $\hat{\lambda}$, where $w_i$ and $\hat{w}_i$ are real numbers (weights) and:

$$\begin{array}{l} \sum_{i=1}^{M} w_i = 1 \quad ; \quad \sum_{i=1}^{M} \hat{w}_i = 1 \\ f_i = N(\mu_i, \Sigma_i) \quad ; \quad \hat{f}_i = N\left(\hat{\mu}_i, \hat{\Sigma}_i\right) \end{array} \tag{4}$$

and whose components $f_i$ and $\hat{f}_j$ are corresponding when $i = j$[1]. We can develop the de nition of the KL divergence between $f$ and $\hat{f}$ in the following way:

$$\begin{aligned} & D\left(f\left|\hat{f}\right.\right) = D\left(\sum_{i=1}^{M} w_i f_i \left| \sum_{i=1}^{M} \hat{w}_i \hat{f}_i \right.\right) \\ = & \int \left(\sum_{i=1}^{M} w_i f_i\right) \log \frac{\sum_{i=1}^{M} w_i f_i}{\sum_{i=1}^{M} \hat{w}_i \hat{f}_i} \leq \int \sum_{i=1}^{M} \left[ w_i f_i \log \frac{w_i f_i}{\hat{w}_i \hat{f}_i} \right] \\ = & \sum_{i=1}^{M} w_i \log \frac{w_i}{\hat{w}_i} + \sum_{i=1}^{M} w_i \int f_i \log \frac{f_i}{\hat{f}_i} \end{aligned} \tag{5}$$

where the inequality proofs using the log-sum inequality [5][8]. Therefore, the KL divergence between the two GMM models is upper-bounded by two components. The rst term is related to the weights of the distribution and can be easily computed. The second term is the sum of the divergences of the individual corresponding Gaussian mixtures, which can be computed using the following formula [5]:

$$\begin{aligned} & \int f_i \log \frac{f_i}{\hat{f}_i} \\ = & \frac{1}{2}\left[ \log \frac{\det(\Sigma_i)}{\det\left(\hat{\Sigma}_i\right)} - \dim(\Sigma_i) + tr\left(\hat{\Sigma}_i^{-1}\Sigma_i\right) \right. \\ + & \left. (\mu_i - \hat{\mu}_i)\,\hat{\Sigma}_i^{-1}\,(\mu_i - \hat{\mu}_i) \right] \end{aligned} \tag{6}$$

[1]The correspondence between these Gaussian components is assumed because in our case both GMM distributions will come from the same UBM via mean-only MAP adaptation

In our case, the two GMMs are adapted from the same model using only-means MAP adaptation. Furthermore, we use diagonal covariance matrices for each mixture component. So, the KL divergence is symmetric, and therefore it is equivalent to a distance. Hence, the computation process in this situation is reduced to the simple expression

$$\int f_i \log \frac{f_i}{\hat{f}_i} = \frac{1}{2}\left[ (\mu_i - \hat{\mu}_i)\,\hat{\Sigma}_i^{-1}\,(\mu_i - \hat{\mu}_i) \right] \tag{7}$$

which is very fast compared with other techniques such as Monte-Carlo estimation. In [5], experiments show that this upper bound is very similar to the Monte-Carlo estimated KL divergence. The aforementioned advantages of this measure in the model domain make it very useful in many areas in which it is necessary to compute distances between models, as it is shown for example in [7] for speaker diarization.

# 3. KL DISTANCE FOR SPEAKER-ADAPTED TNORM

The described KL distance upper bound will be used to select speaker-adapted cohorts for TNorm in speaker veri cation systems. For each target speaker model $\lambda_t$, we compute a set of KL distance upper-bounds $D_{t,I} = \{D(f_t|f_{I,1}), \cdots, D(f_t|f_{I,N})\}$ using Equations 5 and 7. If we select only the $K$-Nearest impostor models to $\lambda_t$ following the KL distance criterion, we will have a set of impostor models $\Lambda_{KL-I} = \{\lambda_{KL-I,1}, \cdots, \lambda_{KL-I,K}\}$ to obtain an estimate of the Gaussian pdf for the impostor scores $S_{KL-I} = \{s(O, \lambda_{KL-I,1}), \cdots, s(O, \lambda_{KL-I,K})\}$:

$$f(x|S_{KL-I}) = N(\mu_{KL-TNorm}, \sigma_{KL-TNorm}) \tag{8}$$

So, TNorm will be performed as:

$$s_{KL-TNorm}(O, \lambda_t) = \frac{s(O, \lambda_t) - \mu_{KL-TNorm}}{\sigma_{KL-TNorm}} \tag{9}$$

For each $s(O, \lambda_t)$ computed in each test segment vs. speaker model trial, the application of KL-TNorm can be summarized in the following steps:

- Computation of $D_{t,I}$ using Equations 5 and 7.
- Selection of $K$ impostor models nearest to $\lambda_t$ and obtention of $S_{KL-I}$.
- Estimate Gaussian impostor score distribution for $S_{KL-I}$ and apply Equation 9.

It is important to remark that KL-TNorm can be applied to a non-GMM speaker recognition system, such as a SVM system. We simply have to compute $D_{t,I}$ using GMM modeling and then use this distance set to select the KL-TNorm cohorts for the non-GMM system.

# 4. EXPERIMENTS

## 4.1. Databases and Experimental Framework.

The experiments described in this section have been performed using the evaluation protocol proposed by NIST in its 2005 Speaker Recognition Evaluation (SRE) [6]. The database used

in this evaluation is a subcorpus of the MIXER database, which has been collected using the sher protocol, and which contains data recorded across different communication channels (landline, GSM, CDMA, etc.), using different handsets and microphones (carbon button, electret, earphones, cordless, etc.) and different languages (American English, Arabic, Spanish, Mandarin, etc.). The evaluation protocol gave the possibility of presenting a system to any task involving one of the following training conditions: 10 seconds, 1, 3 and 8 conversation sides; and one of the following test conditions: 10 seconds, 1 conversation side, 3 full conversations in a mixed channel and multichannel microphone data. Each conversation side had an average duration of 5 minutes, having 2.5 minutes aprox. after silence removal. Although there were speaker of both genders in the corpus, no cross-gender trials were performed. Details can be found in NIST SRE 2005 Evaluation Plan [6]. We present experiments using the ATVS GMM [9] and SVM systems presented to NIST SRE 2005. Before the evaluation, a development set was selected for system tuning. This set consisted of the database used for the NIST SRE 2004 Evaluation, which is also a subset of MIXER. Additional data needed for development trials (UBM, normalization cohorts, etc.; namely background data) was selected from the same development set, and we used the same NIST SRE 2004 database as background data for the evaluation. The experiments have been performed for 1 conversation side testing and both 1 and 8 conversation side testing (1c-1c and 8c-1c respectively). The TNorm cohorts consist of the NIST SRE 2004 target models for each training condition, and the total number of models $N$ in each cohort is shown in Table 1. Target and cohort models conditions regarding gender and amount of training data were matching in all cases.

Table 1: Total number of models $N$ in each TNorm cohort

| | 1c-1c | | 8c-1c | |
|---|---|---|---|---|
| | male | female | male | female |
| $N$ in cohort | 246 | 370 | 170 | 205 |

The results presented below are organized as follows: rst, we present tuning experiments in the development set in order to choose the KL-TNorm parameters for the evaluation. Then, the results in NIST SRE 2005 for both systems are presented. Finally, we show post-evaluation results. All these experiments prove the improvement that KL-TNorm introduces in the systems under the various conditions presented.

## 4.2. Development Experiments

Tables 2 and 3 summarize the experiments peroformed in the development set. The number of models $K$ selected by KL-TNorm has been varied, and results have been compared with those in which TNorm is performed using the whole cohorts. In table 2, it can be observed that KL-TNorm improves the system performance in terms of Equal Error Rate (EER). Also, we note that this improvement is generally stable for the 50-75 range. It is also observed that the optimum EER value is obtained for a number of selected models of $K = 50$. Table 3 shows the optimum Detection Cost Function (DCF) as de ned by NIST evaluations [6] for the same experiments. It can be pointed out that, while a general trend of improvement exists in all cases, the optimum DCF is not signi cantly lower when KL-TNorm is performed.

Table 2: EER for different systems, training conditions and number of selected models using the development set

| Selected Models (K) | GMM | | SVM | |
|---|---|---|---|---|
| | 1c-1c | 8c-1c | 1c-1c | 8c-1c |
| 25 | 12.73 | 7.47 | 17.91 | 15.61 |
| 50 | 12.47 | 7.41 | 17.29 | 14.23 |
| 75 | 12.56 | 7.47 | 17.65 | 14.33 |
| 100 | 12.74 | 7.54 | 17.77 | 14.38 |
| 125 | 12.76 | 7.54 | 18.01 | 15.49 |
| 150 | 12.85 | 7.67 | 18.25 | 15.69 |
| Total (TNorm) | 13.10 | 7.81 | 18.58 | 15.40 |

Table 3: Minimum DCF for different systems, training conditions and number of selected models using the development set

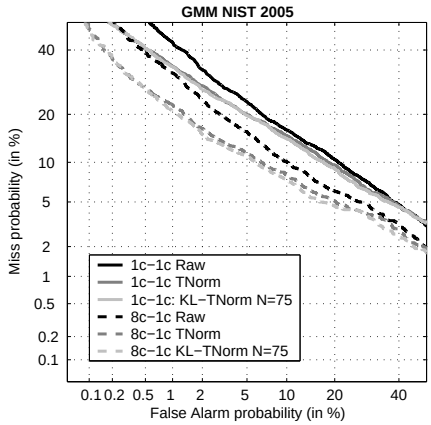
| Selected Models (K) | GMM | | SVM | |
|---|---|---|---|---|
| | 1c-1c | 8c-1c | 1c-1c | 8c-1c |
| 25 | 0.049 | 0.032 | 0.071 | 0.061 |
| 50 | 0.047 | 0.031 | 0.068 | 0.059 |
| 75 | 0.047 | 0.031 | 0.068 | 0.058 |
| 100 | 0.047 | 0.031 | 0.067 | 0.059 |
| 125 | 0.048 | 0.031 | 0.067 | 0.059 |
| 150 | 0.048 | 0.031 | 0.068 | 0.059 |
| Total (TNorm) | 0.048 | 0.031 | 0.077 | 0.059 |

## 4.3. NIST SRE 2005 Results

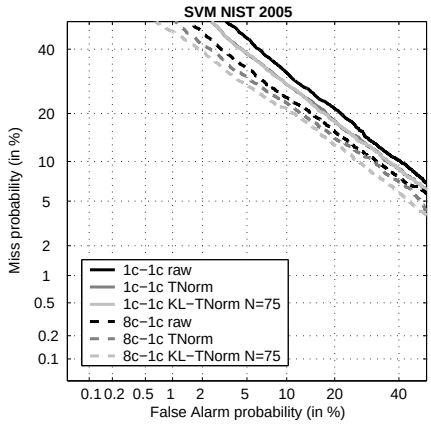
In order to perform KL-TNorm in NIST SRE 2005, we set $K = 75$ for both systems and conditions. For that value the system presents a good performance in development experiments, and the number of models selected is suf ciently high to be robust in blind conditions. Figure 1 show the performance of the GMM and SVM systems in both 1c-1c and 8c-1c conditions when no normalization, TNorm and KL-TNorm are used. For all these cases, TNorm has been applied using the total number of models in the cohorts (Table 1). We note signi cant improvement in system performance for the 8conv4w-1conv4w condition, whereas this is not appreciated for the 1conv4w-1conv4w condition.

## 4.4. Post-Evaluation Experiments

In order to demonstrate the capabilities of the KL-TNorm technique in all evaluation conditions presented, we have compared it to TNorm when the number of models in the cohorts is equal for both normalization schemes. As before, we have used $K = 75$ selected models in KL-TNorm for all experiments. In an analogous way, we have performed TNorm using $K = 75$ models randomly selected from each whole cohort. In order to perform statistically signi cant experiments, we have evaluated the system results using 10 different random selections, and we have averaged the EER and DCF values obtained. The results in Table 4 for the GMM system show signi cant and stable improvement for all EER values in all conditions. On the other hand, there is some improvement in DCF values in all cases, but it is very small. Table 5 shows the same results for the SVM system, which are even better than those for the GMM system, both for EER and DCF values.




(a)




(b)

Figure 1: KL-TNorm in blind NIST SRE 2005 Evaluation. (a) GMM systems and (b) SVM systems.

## 5. CONCLUSIONS

In this paper we have presented a novel technique for score normalization, namely KL-TNorm, which performs speaker-adapted TNorm by selecting the nearest models to each speaker model from a given cohort. This selection is performed via a fast approximation of the Kullback-Leibler divergence for GMM as a measure of similarity between statistic models. The experiments presented in this paper using the NIST SRE 2005 evaluation protocol show signi cant improvement in the system performance when using this technique in all the presented cases. Moreover, it can be observed that the results are stable in an wide range of selected model numbers. The main advantage of the presented technique is based on the fast distance computation algorithm chosen, which allow the selection of sub-cohorts from very big sets of models, and therefore optimizes the test-normalization computational cost. Furthermore, the computed distances from GMM models can be ef ciently used for cohort selection in any other non-GMM speaker recognition system, as it has been demonstrated in the experiments presented using the ATVS SVM system. Therefore, the use of this technique in real applications is highly interesting, as the need of adaptation in some cases where high mismatch between populations and speakers model exists represent an important problem in the use of this kind of normalization methods. Finally, we remark that the usefulness and fast performance of the described Kullback-Leibler divergence approximation makes it a very useful tool for any area in which a measure of distances between models is needed.

## 6. REFERENCES

[1] Auckentaller, R., Carey, M., Lloyd-Thomas, H.: Score Normalization for Text-Independent Speaker Veri cation Systems. Digital Signal Processing, **(**10) (2000) 54-42

[2] Reynolds D.: Comparison of Background Normalization Methods for Text-independent Speaker Veri cation. Proc. Eurospeech (1997)

[3] Reynolds, D., Quatieri, T., Dunn, R.: Speaker Veri cation Using Adapted Gaussian Mixture Models. Digital Signal Processing, **(**10) (2000) 19-41

[4] Sturim, D., Reynolds, D.: Speaker Adaptive Cohort Selection for TNorm in Text-independent Speaker Veri cation. Proc. ICASSP (2005) 741-744

[5] Do, M. N.: Fast Aproximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models. Signal Processing letters, **(**10) (2003) 115-118

[6] NIST Speech Group: http://www.nist.gov/speech

[7] Ben, M. et al.: Speaker Diarization using bottom-up clustering based on a Parameter-derived Distance between adapted GMMs. Proc. ICSLP (2004)

[8] Cover, T. M., Thomas, J. A.: Elements of Information Theory. Wiley Interscience (1991)

[9] Gonzalez-Rodriguez, J. et al.: On Robust Estimation of Likelihood Ratios: The ATVS-UPM System at 2003 NFI/TNO Forensic Evaluation. Proc. ODYSSEY (2004) 83-90

Table 4: TNorm and KL-TNorm in GMM system using $K = 75$ selected models for both techniques (TNorm values are averaged in 10 random selection trials)

| $GMM$ $K = 75$ | 1c-1c | | 8c-1c | |
|---|---|---|---|---|
| | male | female | male | female |
| EER TNorm (Av.) | 11.14 | 14.62 | 7.78 | 9.57 |
| EER KL-TNorm | 10.76 | 13.88 | 7.25 | 9.12 |
| EER Av. Improvement | 3.4% | 5.0% | 6.8% | 4.7% |
| DCF TNorm (Av.) | 0.041 | 0.048 | 0.030 | 0.033 |
| DCF KL-TNorm | 0.039 | 0.047 | 0.029 | 0.031 |

Table 5: TNorm and KL-TNorm in GMM system using $K = 75$ selected models for both techniques (TNorm values are averaged in 10 random selection trials)

| $SVM$ $K = 75$ | 1c-1c | | 8c-1c | |
|---|---|---|---|---|
| | male | female | male | female |
| EER TNorm (Av.) | 19.22 | 19.27 | 16.58 | 16.79 |
| EER KL-TNorm | 17.19 | 17.87 | 14.15 | 14.59 |
| EER Av. Improvement | 11.6% | 7.3% | 14.7% | 13.1% |
| DCF TNorm (Av.) | 0.073 | 0.075 | 0.060 | 0.060 |
| DCF KL-TNorm | 0.063 | 0.073 | 0.057 | 0.059 |

# MULTIMODALITY AND EMERGING TECHNOLOGIES

# A MATCHING-SCORE NORMALIZATION TECHNIQUE FOR MULTIMODAL BIOMETRIC SYSTEMS

*Slobodan Ribaric and Ivan Fratric*

Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, 10000 Zagreb, Croatia
slobodan.ribaric@fer.hr, ivan.fratric@fer.hr

## ABSTRACT

In this paper we propose a new technique for the matching-score normalization for multimodal biometric systems. The technique is based on a piece-wise linear function, the parameters of which are obtained automatically from genuine and impostor matching-score distributions that are generated on a training data. The technique is compared to other score normalization techniques for a bimodal verification system based on the principal lines of a palmprint and eigenfaces.

## 1. INTRODUCTION

A multimodal biometric system [1] requires an integration scheme to fuse the information obtained from the individual modalities. There are various levels of fusion [2, 3]:

i) Fusion at the feature-extraction level, where the features extracted using two or more sensors are concatenated.

ii) Fusion at the matching-score level, where the matching scores obtained from multiple matchers are combined.

iii) Fusion at the decision level, where the accept/reject decisions of multiple systems are consolidated.

Fusion at the matching-score level is the most popular and frequently used method because of its good performance, intuitiveness and simplicity [4, 5].

When using the method of fusion at the matching-score level a normalization step is generally required for the following reasons [5]:

i) The matching scores at the output of the individual matchers can be represented in different ways. For example, one matcher may output distances (as a measure of dissimilarity), while the others may output proximities (as a measure of similarity).

ii) The matcher outputs can be in different numerical ranges.

iii) The genuine and impostor matching scores from different modalities might not follow the same statistical distributions.

The above reasons illustrate how normalization at the matching-score level is a critical point in the design of a multimodal biometric identification or verification system [5, 6].

The rest of this paper is organized as follows. Section 2 briefly describes some of the frequently used normalization techniques for multimodal biometric systems, as well as a new technique for normalization based on a piece-wise linear function. In Section 3, the prototype of a bimodal biometric system based on the fusion of the principal lines of the palmprint and eigenface features is described. The system has been used to test different normalization techniques. The experimental results of the person verification, based on the testing of different normalization techniques, are presented in Section 4. Some concluding remarks are given in Section 5.

## 2. NORMALIZATION TECHNIQUES

In this section we present some of the well-known normalization techniques used in multimodal biometric systems. A novel technique based on the piece-wise linear function is also described.

In general, the normalized score is obtained by using a normalization function. Let $r_i \in O_i$, $i = 1, 2, ..., M$, where $r_i$ is the output raw value of the matcher $i$ ($M$ is the total number of matchers). $O_i$ is the set of all the raw output values of the corresponding matcher $i$ (matching scores). Usually, $O_i \subset \mathfrak{R}^1$. The corresponding normalized matching scores are denoted as $n_i \in N_i$, where $N_i \subset \mathfrak{R}^1$ is the set of normalized matching scores of the matcher $i$. The normalization is defined as a function that maps $O_i$ to $N_i$. During training, the set $O_i$ can be divided into two subsets, $O_i^G$ and $O_i^I$, which denote the genuine and imposter raw matching scores, respectively.

**Min-max normalization**

The *min-max normalization* is the simplest normalization technique that achieves the common numerical range of the scores ($N_i = [0, 1]$) and also retains the shapes of the original distributions. As with most normalization techniques, there are two cases relating to the character of the raw scores:

i) Similarity scores

ii) Distance scores

For the min-max normalization we will illustrate both the similarity scores and the distance score normalization.

The min-max normalization function is given as:

a) For the raw similarity scores $r_i$

$$n_i = \frac{r_i - \min(O_i)}{\max(O_i) - \min(O_i)} \qquad (1a)$$

b) For the raw distance scores $r_i$

$$n_i = \frac{\max(O_i) - r_i}{\max(O_i) - \min(O_i)} \qquad (1b)$$

In general, the values $\max(O_i)$ and $\min(O_i)$ are obtained from the training set.

In the rest of the paper we will present the similarity scores normalization only. The distance scores normalization functions can be obtained analogously.

**Z-score normalization**

The *Z-score normalization* retains the shape of the input distributions as the output only if they are Gaussian-distributed, and it does not guarantee a common numerical range. The similarity scores normalization is given as

$$n_i = \frac{r_i - \mu}{\sigma} \qquad (2)$$

where $\mu$ and $\sigma$ are the mean and the standard deviation of the matching score set $O_i$, respectively.

**Median-MAD normalization**

The *median-MAD* (median absolute deviation) *normalization* does not guarantee the common numerical range and is insensitive to outliers. The normalization is given as

$$n_i = \frac{r_i - median}{MAD} \qquad (3)$$

where *median* is the median of the values in $O_i$ and $MAD = median(|r_i - median|)$.

**Double-sigmoid normalization**

The *double-sigmoid normalization* [7] is given as:

$$n_i = \begin{cases} \dfrac{1}{1+\exp\left(-2\left(\dfrac{r_i - t}{l_1}\right)\right)} & if\ r_i < t \\ \dfrac{1}{1+\exp\left(-2\left(\dfrac{r_i - t}{l_2}\right)\right)} & otherwise \end{cases} \qquad (4)$$

where $t$, $l_1$ and $l_2$ are the parameters that control the shape of the normalization function on the different segments. We have chosen the parameters so that $t$ is the mean value of the $\min(O_i^G)$ and $\max(O_i^I)$. The parameters $l_1$ and $l_2$ are chosen as $l_1 = t - \min(O_i^G)$ and $l_2 = \max(O_i^I) - t$.

**Tanh-normalization**

The normalization based on the *tanh-estimators* [8] is reported to be robust and highly efficient [5]. It is given as:

$$n_i = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{r_i - \mu}{\sigma}\right)\right)\right\} \qquad (5)$$

where μ and σ denote the mean and standard deviation of the $O_i^G$, respectively. According to [5] the results of this normalization technique are quite similar to those produced by the Z-score normalization.

Instead of the real values of μ and σ, their estimated values based on the Hampel estimators can be used in order to make the normalization technique more robust. However, we have observed that for a training set not containing artificially introduced outliers, the use of Hampel estimators gives a nearly identical multimodal system performance as when using the real values of μ and σ.

**Piecewise-linear normalization – a new normalization technique**

We propose a new normalization technique that results in a common numerical range of [0, 1] and strives to achieve good separation of the genuine and the impostor matching-score distributions. We construct the normalization function heuristically, so that it has a constant value of 1 for the non-overlapping area of the genuine matching-score distribution, and has a constant value of 0 for the non-overlapping area of the impostor matching-score distributions. In between, i.e., in the overlapping area of the impostor and genuine matching-score distributions, the function changes linearly.

The normalization function is a three-segmented piecewise-linear function:

a) For the raw similarity scores $r_i$

$$n_i = \begin{cases} 0 & if\ r_i < \min(O_i^G) \\ \dfrac{r_i - \min(O_i^G)}{\max(O_i^I) - \min(O_i^G)} & if\ \min(O_i^G) < r_i < \max(O_i^I) \\ 1 & if\ r_i > \max(O_i^I) \end{cases} \qquad (6a)$$

b) For the raw distance scores $r_i$

$$n_i = \begin{cases} 1 & if\ r_i < \min(O_i^I) \\ \dfrac{\max(O_i^G) - r_i}{\max(O_i^G) - \min(O_i^I)} & if\ \min(O_i^I) < r_i < \max(O_i^G) \\ 0 & if\ r_i > \max(O_i^G) \end{cases} \qquad (6b)$$

The piecewise-linear normalization function for distances in an example of genuine and impostor score distributions can be seen in Figure 1.

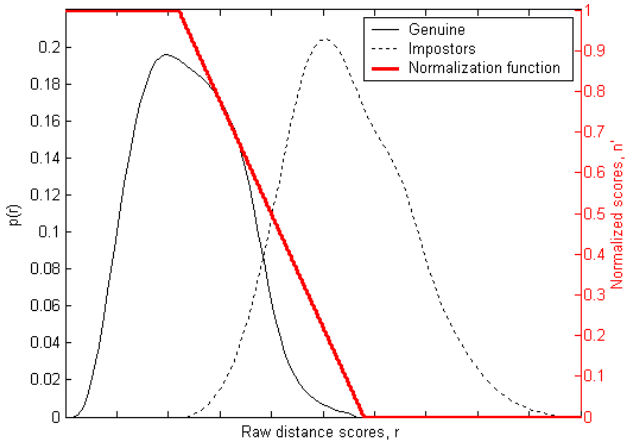



**Figure 1**: Example of genuine and impostor score distributions with the matching-score normalization function for distances

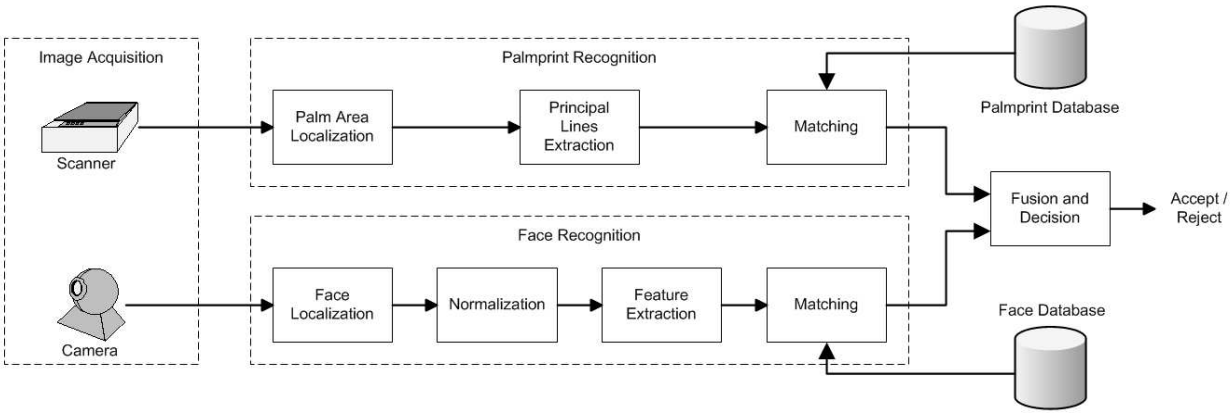



**Figure 2**: Block-diagram of the bimodal biometric verification system based on palmprint and facial features

## 3. A BIMODAL BIOMETRIC VERIFICATION SYSTEM BASED ON PALMPRINT AND FACIAL FEATURES

The described normalization techniques have been tested on a multimodal biometric verification system based on palmprint and facial features [9]. The block-diagram of the system is shown in Figure 2.

The processing of the palmprint and face images, up to the point of fusion, is carried out separately in the palmprint-recognition and the face-recognition subsystems. In the first phase of the palmprint-recognition process the area of the palm is located on the basis of the hand contour and the stable points. In the second phase, the principal lines of the palm are extracted using line-detection masks and a line-tracking algorithm. Finally, a live-template based on the principal palm lines is matched to the templates from the palmprint database using an approach similar to the HYPER [10] method. The result of the palm-line template matching is the similarity measure $Q$. Figure 3 shows the similarity measure $Q$ for several palm-line template pairs.

The process of face recognition consists of four phases: face localization, based on the Hough method [11]; normalization, including geometry and lighting normalization; feature extraction using eigenfaces [12]; and finally, matching of the live-template to the templates stored in the face database. The result of the eigenface template matching is the Euclidean distance $d$.

The normalized matching scores from both recognition modules are combined into a unique matching score using sum fusion at the matching-score level. Based on this unique matching score, a decision about whether to accept or reject a user is made using thresholding.

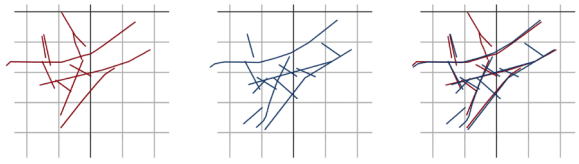
## 4. EXPERIMENTAL RESULTS

To evaluate the performance of the system, based on testing the previously described normalization techniques, a database containing palm and face samples was required. The XM2VTS frontal-face-images database [13] was used as the face database. We collected the hand database ourselves using a scanner. The spatial resolution of the hand images is 180 dots per inch (dpi) / 256 grey levels. A "chimerical" multimodal database was created using pairs of artificially matched palm and face samples.

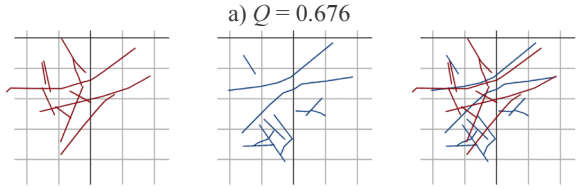
The database was divided into two sets: the training set and the testing set. The training set consisted of 440 image pairs of 110 people (4 image pairs per person) and was used as a training database for the individual modalities, as well as to get the distributions of the unimodal matching scores used in the learning of the parameters for the normalization techniques.

The testing dataset consisted of 1048 image pairs of 131 people (8 image pairs per person) and was used exclusively for the evaluation of the system performance. Out of 8 image pairs for each person, 5 were used in the enrolment stage and 3 were used for testing. The tests involved trying to verify every test pair for every one of the 131 people enrolled in the database. This setup makes for 393 (131 x 3) valid-client and 51,090 (131 x 3 x 130) impostor experiments.

Figure 4 shows the ROC obtained from the bimodal biometric verification system using different normalization techniques. Also, in Table 1, the EER (equal error rate) and the minimum TER (total error rate) values obtained using different normalization techniques are given.



a) $Q = 0.676$



b) $Q = 0.071$

**Figure 3**: Comparison of palmprint templates and the similarity measure $Q$: a) – comparison of palmprint templates of the same person; b) – comparison of palmprint templates of different people. The first and the second columns represent the individual palmprint templates. The third column represents both templates in the same coordinate system

As can be seen from the results, the new piecewise-linear normalization gives the best results in terms of the EER and the minimum TER, followed closely by the median-MAD normalization (Table 1). Also, when comparing our experimental results to those obtained by Jain et al. [5] using all of the described normalization techniques (except the piecewise-linear normalization), but on a different biometric system (using fingerprint, facial and hand-geometry features), some differences can be observed. For example, the median-MAD normalization, which performs poorly on the system used by Jain et al., gives exceptionally good results for our system. On the other hand, some of the techniques that give good results in the system used by Jain et al. do not perform so well with our system.

| Normalization technique | piecewise-linear | min-max | z-score | median-MAD | sigmoid | tanh |
|---|---|---|---|---|---|---|
| EER | 2,79% | 3,12% | 3,15% | 2,79% | 3,81% | 3,05% |
| min TER | 5,15% | 6,39% | 5,56% | 5,42% | 5,72% | 5,74% |

**Table 1**: EER and the minimum TER values for the different normalization techniques obtained on a bimodal verification system

## 5. CONCLUSION

The experiments showed that the relatively simple piecewise-linear normalization technique gives the best results for our multimodal system and the test database.

Comparing our results with those obtained by Jain et al. [5] leads us to conclude that there is no single normalization technique, among those tested, which would perform best for all multimodal biometric systems. The performance of the normalization methods would, most likely, be dependent on the shape of the distribution of the matching scores in the multimodal biometric system itself. Exhaustive testing of the different normalization techniques should be performed in order to select the best-performing technique for the particular system under investigation.

## 6. REFERENCES


1. L. Hong and A.K. Jain, "Multimodal Biometrics", in A.K. Jain, R. Bolle and S. Pankanti (Eds.): Biometrics: Personal Identification in a Networked Society, Kluwer Academic Publishers, 1999, pp. 326-344

2. A. Ross and A. Jain, "Information Fusion in Biometrics", Pattern Recognition Letters, vol. 24, 2003, pp. 2115-2125

3. J. Kittler and F.M. Alkoot, "Sum Versus Vote Fusion in Multiple Classifier Systems", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 1, 2003, pp. 110-115.

4. M. Indovina, U. Uludag, R. Snelick, A. Mink and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach", *Proc. MMUA 2003, Workshop on Multimodal User Authentication*, Santa Barbara, CA, December 11-12, 2003, pp. 99-106

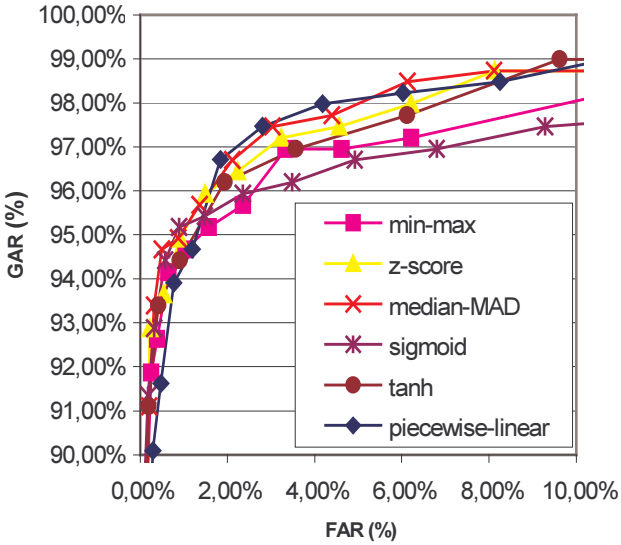




**Figure 4**: ROC curves obtained from the bimodal biometric verification system using different normalization techniques


5. A. K. Jain, K. Nandakumar and A. Ross, "Score Normalization in Multimodal Biometric Systems", *to appear in Pattern Recognition*, 2005.

6. R. Auckenthaler, M.J. Carey and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", Digital Signal Processing, vol. 10, nos. 1-3, 2000, pp. 42-54

7. R. Cappelli, D. Maio and D. Maltoni, "Combining fingerprint classifiers", in Proc. of First International Workshop on Multiple Classifier Systems, 2000, pp. 351–361.

8. F.R. Hampel, P.J. Rousseeuw, E.M. Ronchetti and W.A. Stahel, "Robust Statistics: The Approach Based on Influence Functions", Wiley, New York, 1986.

9. S. Ribaric, I. Fratric, K. Kis, "A Biometric Verification System Based on the Fusion of Palmprint and Face Features", to be published in Proc. of 4th International Symposium on Image and Signal Processing and Analysis (ISPA 2005), September 15-17, 2005, Zagreb, Croatia

10. N. Ayache and O. D. Faugeras, "A New Approach for the Recognition and Positioning of Two-Dimensional Objects", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 8, no. 1, 1986, pp. 44-54.

11. R. Jain, R. Kasturi, B. G. Schunk, Machine Vision, McGraw-Hill, Inc., New York, 1995.

12. M. Turk and A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience, vol. 3, no. 1, 1991, pp. 71-86.

13. K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre, "XM2VTSDB: The Extended M2VTS Database", Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99), Washington D.C., 1999, pp. 72-77.

# MYIDEA – MULTIMODAL BIOMETRICS DATABASE, DESCRIPTION OF ACQUISITION PROTOCOLS

*Bruno Dumas[1], Catherine Pugin[1], Jean Hennebert[1], Dijana Petrovska-Delacrétaz[2], Andreas Humm[1], Florian Evéquoz[1], Rolf Ingold[1], and Didier Von Rotz[3]*

[1] DIVA Group, Department of Informatics, University of Fribourg
Ch. Du Musée 3, 1700 Fribourg, Switzerland
[2] INT, Dept. EPH, Intermedia, 9 rue Charles Fourier 91011 Evry, France
[3] EIF, Bd de Pérolles 80 – CP 32, 1705 Fribourg, Switzerland

## ABSTRACT

This document describes the acquisition protocols of MyIDea, a new large and realistic multimodal biometric database designed to conduct research experiments in *Identity Verification (IV)*. The key points of MyIDea are threefold: (1) it is strongly multimodal; (2) it implements realistic scenarios in an open-set framework; (3) it uses sensors of different quality to record most of the modalities. The combination of these three points makes MyIDea novel and pretty unique in comparison to existing databases. Furthermore, special care is put in the design of the acquisition procedures to allow MyIDea to complement existing databases such as BANCA, MCYT or BIOMET. MyIDea includes talking face, audio, fingerprints, signature, handwriting and hand geometry. MyIDea will be available early 2006 with an initial set of 104 subjects recorded over three sessions. Other recording sets will be potentially planned in 2006.

## 1. OVERVIEW

Multimodal biometrics has raised a growing interest in the industrial and scientific community. The potential increase of accuracy combined with better robustness against forgeries makes indeed multimodal biometrics a promising field. Moreover, important issues like the correlation of the modalities or the combination of multiple and potentially heterogeneous sources of biometric features make the field attractive from a research point of view. However, there is a clear need for multimodal databases, given that current existing databases offer few recorded modalities, often implement unrealistic closed-set scenarios and sometimes use even more unrealistic acquisition conditions.

MyIDea aims at providing the scientific community with a multimodal database including talking face, voice, fingerprints, signature, handwriting, palmprint and hand geometry. Real-life multimodal systems will probably not use all of these modalities at the same time. However, MyIDea will allow to group modalities according to some scenario such as, for example, biometric passports including face and fingerprint. Grouping modalities can also be performed according to the capacities of next-generation sensors that will be able to simultaneously acquire multiple modalities (hand-fingerprint, voice-handwriting, etc).

The specifications of MyIDea can be summarized as follows:

1. target of 104 subjects, recorded over three different sessions spaced in time; no control of minimum or maximum interval between sessions as in real-life applications
2. direct compatibility and potential extension of existing mono or multimodal databases: BANCA [1], BIOMET [2], XM2VTSDB [3], MCYT [4] and IAM [5]
3. different types and qualities of sensors, various and realistic acquisition scenarios, content recorded in French and English
4. organization of the recordings to allow for open-set experiments
5. impostor attempts for voice, signature and handwriting

MyIDea also offers a less-common, novel acquisition mode which is the bimodal voice-signature and voice-handwriting. It has been experimented that these modalities can be simultaneously recorded in a scenario where the user is asked to utter what he is writing. Robustness against impostor attacks is the main advantage of this bimodal approach. The act of speaking and writing at the same time is indeed perceived as a rather easy cognitive task by the user. However, it can be reasonably expected that impostors will have some difficulties to imitate the writing or the signature of someone else while speaking. Another novel point about the signature modality is the use of dedicated software which renders the dynamics of the signal to help impostors to imitate the signature.

At the time of writing this article, specific evaluation protocols were not yet defined on MyIDea. However, most of the parts of MyIDea are compatible with existing databases for which evaluation protocols are already available. Therefore, one can already use these protocols and apply them to MyIDea data. Another possibility is to use MyIDea to extend the evaluation sets defined in these protocols in order to obtain more significant results.

The overall MyIDea project is performed in the framework of collaborations between the University of Fribourg in Switzerland [6], the Engineering School of Fribourg in

Switzerland [7] and the GET in Paris [8]. In Fribourg, MyIDea is supported by the Swiss project IM2 [9]. Feedback and interest have also been raised from the EC Network of Excellence BioSecure [10].

This document is a summarized version of a detailed research report [11] available from the MyIDea website [12]. The organization is as follows: section 2 and 3 give an overview of the acquisition system and of the acquisition protocols; these sections share the same structure which is organized by modality; section 4 introduces briefly the software that has been built to perform the recordings; finally section 5 outlines the availability of the database and future work.

# 2. ACQUISITION SYSTEM

## 2.1 Video for voice and face

The settings used to record the voice and face of subjects were inspired by BANCA, BIOMET and XM2VTS. The hardware used to record the videos is of two types: on one side, "good quality" Sony Digital Camera *DCR-HC40* recording in DV quality; on the other side, regular webcams. The audio is captured by high-quality Shure microphones and directly by the integrated microphone of the webcams when applicable. This setting has been reproduced for three different acquisition scenarios similar to the one of BANCA: controlled, degraded and adverse.

In the controlled scenario, people sit in front of a uniform blue wall at a fixed distance of the camera. Lighting conditions are verified: three halogens are placed in order to avoid shadows on the face of the subjects; mean luminance value on the surface of the face is controlled around 650 lux; blinds of the room are shut. The automatic setups of the camera are turned off to avoid unwanted variations in the recordings. Both good quality and web cameras are used in this scenario.

In the degraded scenario, a work environment is simulated. The subject is sitting at a desk, reading data from a computer screen on top of which are placed the cameras. The recording conditions correspond to those of a workplace with a mix of natural light and office lamps, with background noise coming from both a nearby road and with people potentially present in the background. The default automatic settings of the cameras are turned on. Both good quality and web cameras are used for this scenario.

In the adverse scenario, the environment is highly variable, with very little control over the lighting conditions. Background noise is present due to people walking or speaking in the vicinity. The background may change from one session to another, with people passing behind (and sometimes even in front of) the subject. The uncontrolled events are sometimes distracting the subject who shows unwanted movements or changes in intonation. Only the good quality camera is used in this scenario.

## 2.2 Fingerprints

Two fingerprint sensors are used for the acquisition. Both sensors are controlled by the software developed for the acquisition campaign (see section 4.2) and the drivers provided by the constructor are used.

The first sensor is a high-end optical sensor *Morphosmart MSO-100* from SAGEM [13]. Fingerprints are acquired at 500 dpi on a 21 mm x 21 mm acquisition area, 8-bit grey scale. This reference sensor has been chosen to get direct compatibility with the BIOMET database.

The second sensor is the *TocaBit* scan-thermal sensor from Ekey [14] which is based on a *FCD4B14 FingerChip* from Atmel [15]. This sensor measures the temperature differential between the sensor pixels that are in contact with the finger ridges and those that are not. Provided that the fingertip has been swept at an appropriate speed and pressure, the overlap between successive frames enables an image of the entire fingerprint to be reconstructed [16]. The obtained image is typically 25 mm x 14 mm (equivalent to 500 x 280 pixels), at 8 bit grey scale resolution. This sensor presents the advantage that no cleaning is needed between acquisitions. This device has been chosen as, to our knowledge, no database uses this type of scan-thermal procedure. Furthermore, its small size and robustness make it a good candidate to be embedded on devices such as mobile phones.

## 2.3 Palmprints

A *perfection 1660* scanner from EPSON is used to scan palm images of the hand. The scanner is driven by the software developed for the acquisition campaign through the TWAIN driver (see section 4.2). The scanning resolution can be set to various values. As for the optical fingerprint sensor, the glass plate of the scanner needs frequent cleaning to avoid additional noise accumulating between acquisitions. Cleaning is performed with alcohol-soaked napkins at a frequency defined by the protocol.

## 2.4 Hand geometry

Hand geometry including the dorsum surface (top-view) and the lateral surfaces (left and right side-views) of the hand is captured using an *Olympus C-370* CCD camera. A platform consisting of a plane surface and two inclined lateral mirrors has been built. Pegs are placed on the plane surface to guide the position of the user's hand. One shot takes a picture including the top view and the side views of the hand.

## 2.5 Signature and handwriting

Signatures and handwritings are acquired with an *A4 Intuos2* graphic tablet from WACOM [17]. An *Intuos InkPen* is used to write on standard paper positioned on the tablet. This procedure presents the advantages to record on-line and off-line data in the same time and to allow a very natural writing using an almost standard pen and paper. This tablet has been chosen for direct compatibility with the BIOMET, MCYT and IAM databases

For the on-line data, the tablet records 5 parameters: x-y coordinates, pressure, azimuth and altitude at a frequency of 100 Hz. The tablet is managed by the software developed for the acquisition campaign (see section 4.2) using the drivers provided by the constructor. The software also allows performing a synchronized acquisition of writing and voice data when the subject is asked to read aloud what he/she is

writing (see section 3.5 and 3.6). For this purpose, a computer microphone mounted on a headset is used.

# 3. ACQUISITION PROTOCOL

The strategy for the constitution of the groups is similar to the one of the BANCA database [1]. Each group has 13 subjects and is gender specific. Impostures are performed within each group, i.e. there are no cross-group impostures. The goal is to reach 8 groups of 13 subjects, i.e. a total of 104 subjects over 3 sessions. The time-interval between two sessions is not controlled and can range from days to months. The protocol is detailed for each modality in the following sub-sections.

## 3.1 Video for voice and face

Two types of contents are recorded. The first one is similar to the content of the BANCA database [1] and the second one is similar to the content recorded in the BIOMET database [2]:

1. BANCA: This content is recorded in controlled, degraded and adverse scenarios such as described earlier (see section 2.1). The subject is asked to read 10 random digits, a full mail address and a birth date. The subject is then asked to read the content (digits, address, birth date) of another subject of her/his group.
2. BIOMET: This content is recorded using the controlled acquisition scenario. The subject is asked to read digits from 0 to 9, digits from 9 to 0, two phonetically balanced sentences, “yes” and “no”, 10 fixed phrases for text-dependent system, 8 password-like phrases (4 genuine and 4 impostor phrases) and 5 random short phrases for text-independent system evaluation. The BIOMET session ends with head rotation movements for 3D reconstruction algorithms.

## 3.2 Fingerprints

Overall, 6 fingerprint images of all 10 fingers are captured per session. For each finger, 2 images are taken on the optical Sagem sensor in an “uncontrolled” manner, then 2 more images on the same sensor in a “controlled” manner, and finally 2 images on the scan-thermal Ekey sensor.

For the optical sensor, the uncontrolled mode is performed without controlling the quality and centering of the acquired image and without removing previous fingerprint traces that could remain on the sensor glass. The controlled mode is performed by visually controlling the quality and centering of the acquired image and by cleaning the glass plate between each acquisition. It has been observed that too dry fingers lead to degraded images using this sensor. Therefore, for the controlled mode, subjects are asked to wipe their finger on their forehead prior to each acquisition in order to smooth out the dryness.

The scan-thermal sensor automatically discards improper images according to its internal criterion. Acquisitions are performed until a fingerprint is declared valid by the sensor.

## 3.3 Palmprint

Per subject and session, palmprints of the right hand are acquired 4 times: 3 times at a resolution of 150 dpi and one last time at a resolution of 400 dpi. Subjects are asked to spread their fingers naturally and keep their rings if any.

## 3.4 Hand geometry

Three pictures of the dorsal and lateral sides of the right hand are taken per subject and session. These pictures are taken at a nominal resolution of 2048 x 1536 pixels. The macro mode is enabled on the CCD camera in order to enhance the quality of the image. The subject is asked to place his right hand according to the pegs. Rings are not removed from the fingers.

## 3.5 Signature, signature and voice

Per subject and session, 6 signature samples are acquired according to 5 different scenarios:

1. **true signature**: the subject signs using her/his own signature as naturally as possible.
2. **true signature with voice**: the subject is asked to synchronously sign and utter the content of her/his signature. If the signature contains flourish or not readable signs, the user is simply asked to utter her/his name.
3. **impostor static signature**: for each session, the subject is asked to imitate the signature of another subject of the same group. Few minutes of training are given to the subject.
4. **impostor dynamic signature**: imposture on the same signature as in 3, but this time, the subject is allowed to study the dynamics of the signature thanks to a dedicated software (see section 4.2).
5. **impostor static signature with voice**: for each session, the subject is asked to imitate the signature of another subject (different from imposture in 3 and 4) of the same group and to synchronously utter the content of that signature.

## 3.6 Handwriting with voice

The subject is asked to synchronously write and utter the content of a text. The content is composed of a fixed generic phrase containing all the letters of the alphabet, followed by a text of 50 to 100 words, randomly chosen from a corpus. The fixed phrase can be used to evaluate text-dependent systems and the random text can be used to evaluate text-independent systems. First, the subject is asked to write and utter the content of the text with his own handwriting, as naturally as possible. Then (s)he is asked to imitate the handwriting of another member of her/his group. For this last part, a training time limited to few minutes is given to the subject.

# 4. SOFTWARE

## 4.1 Protocol Generation

A software has been built to automatically generate the protocols. The main idea is to limit the impact of human

mistakes by generating a documented protocol for each subject and each session. These generated documents are then printed before each acquisition.

For every subject and every session, there are two types of generated documents. The first one is a protocol to help the assistant monitor the acquisition session. The assistant is supposed to check all points of the protocol before starting with the acquisition of each modality. The second part of the generated documents includes the contents that the subject needs to read or write down. Signature and handwriting are performed on these documents.

## 4.2 Biometrics acquisition

Two dedicated softwares have been developed to help the acquisition: *Biblios* and *SignReplay*.

The first one, Biblios, is used to monitor the acquisition and to control the different sensors. Biblios has been initially developed for the recording of BIOMET and extended here for MyIDea. Biblios guides the assistant through all the steps involved in a biometric data acquisition session. For all modalities, an on-screen live playback helps the assistant verify the quality of the acquired data. Biblios is able to monitor the following devices: scanner, graphic tablet, audio microphone and fingerprint sensors. Video sequences and hand geometry are captured independently of Biblios.

The second one, SignReplay is used to playback the on-line signature data of a given subject on the computer screen. The software plots the (x, y) coordinates as a function of time, while letting the user choose a playback speed, which can be equal or lower than the real speed. This software is used to train impostors to imitate the signature based on the dynamics of the signal. With the help of SignReplay, the user can easily visualize the track followed by the pen, with accelerations and decelerations. SignReplay is also used to verify the integrity and the quality of acquired written data.

# 5. AVAILABILITY AND FUTURE WORK

MyIDea database will be distributed almost freely to research institutions. A fee will be asked to cover the reproduction costs (manpower and hardware). Institutions willing to get the database will have to sign a license agreement presently under construction. The main points of the license agreement will be as follows: (1) data must be used for research purposes only; (2) licensee has no rights to distribute the data to third parties.

Future works include the completion of the acquisition of the 104 subjects (end of 2005), the validation of all data (early 2006) and the definition of standard assessment protocols.

# 6. CONCLUSION

We have presented the acquisition protocols of MyIDea, a new biometrics multimodal database. MyIDea is strongly multimodal including talking-face, audio, fingerprints, signature, handwriting and hand geometry. MyIDea presents interesting features such as compatibility with existing databases, use of multiple sensors for the acquisitions and implementation of realistic scenarios in an open-set framework.

# 7. REFERENCES


[1] E. Bailly-Baillière, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran. The BANCA database and evaluation protocol. In *4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2003.

[2] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux les Jardins, J. Lunter, Y. Ni, and D. Petrovska-Delacrétaz. Biomet: a multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In UK University of Surrey, Guildford, editor, *4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, 2003.

[3] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. In *Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*, 1999.

[4] J. Ortega-Garcia, et al. MCYT baseline corpus: a bimodal biometric database. *IEE Proc.-Vis. Image Signal Process*, 150(6): 395–401, December 2003.

[5] U.-V. Marti and H. Bunke. A full English sentence database for off-line handwriting recognition. In *Proc. of the 5th Int. Conf. on Document Analysis and Recognition (ICDAR'99)*, pages 705–708, 1999.

[6] Fribourg University. http://www.unifr.ch.

[7] Ecole d'Ingénieurs et d'architectes de Fribourg. http://www.eif.ch.

[8] Groupe des Ecoles des Télécommunications, Paris. http://www.get-telecom.fr.

[9] Swiss National Center of Competence in Research on Interactive Multimodal Information Management. http://www.im2.ch.

[10] BioSecure NoE. http://www.biosecure.info.

[11] B. Dumas, F. Evequoz, J. Hennebert, A. Humm, R. Ingold, D. Petrovska, C. Pugin and D. Von Rotz. MyIDea, Sensors Specifications and Acquisition Protocol. University of Fribourg Internal Publication, 05-12, June 2005.

[12] MyIDea Database. http://diuf.unifr.ch/go/myidea/

[13] Sagem S.A. http://www.sagem.com.

[14] Ekey biometric systems GmbH. http://www.ekey.net/.

[15] Atmel Corporation. http://www.atmel-grenoble.com.

[16] Peter Bishop. Atmel's fingerchip technology for biometric security. Technical report, Atmel Corporation, 2002.

[17] Wacom Technology Co. http://www.wacom.com.

# FUSION OF CROSS STREAM INFORMATION IN SPEAKER VERIFICATION

*F. Alsaade, A. Malegaonkar and A. Ariyaeeinia*

{F.Alsaade, A.Malegaonkar, A.M.Ariyaeeinia}@.herts.ac.uk

## ABSTRACT

This paper addresses the performance of various statistical data fusion techniques for combining the complementary score information in speaker verification. The complementary verification scores are based on the static and delta cepstral features. Both LPCC (Linear prediction-based cepstral coefficients) and MFCC (mel-frequency cepstral coefficients) are considered in the study. The experiments conducted using a GMM-based speaker verification system, provides valuable information on the relative effectiveness of different fusion methods applied at the score level. It is also demonstrated that a higher speaker discrimination capability can be achieved by applying the fusion at the score level rather than at the feature level.

## 1. INTRODUCTION

The fusion of the complementary information obtained from the biometric data has been a research area of considerable interest. The efforts in this area are mainly focussed on fusing the information obtained using various independent modalities. For instance, a popular approach is to combine face and voice modalities to achieve a better recognition of individuals. The motivation behind this approach is that the independent information obtained using different modalities is thought to possess complementary evidences about the identity attributes of a particular person [1]. Hence combining such complementary information should be more beneficial than using a single modality. Various statistical fusion techniques have been developed for this task [2]. These range from using different weighting schemes that assign weights to the information streams according to their information content, to support vector machines which use the principle of obtaining the best possible boundary for classification, according to the training data.

Speaker verification is the task of matching the information obtained from a given test utterance against the model associated with the claimed identity. The process involves a binary decision depending on whether or not the match score exceeds a preset threshold. It is therefore desired that the metric adopted for this purpose can effectively discriminate between each true claimant and impostors. The most common approach to representing the registered speaker information is through training the Gaussian Mixture Models (GMM) on the speech feature data [3]. In GMM-based speaker verification, likelihood scores are used as matching metrics. Most of the verification systems use cepstral features to represent the speaker information. Static and delta cepstra obtained from speech represent two distinctive aspects of human vocal tract. Static cepstra represent the coarse aspects of vocal tract configuration under the assumption of being stationary, while delta coefficients represent the time varying (dynamic) information such as speaking style, and speaking rate [4]. This information can be derived from cepstra based on the linear prediction analysis (LPCC), or based on the perceptual processing on filter bank analysis (MFCC). Though delta coefficients are derived from static coefficients using a polynomial fit method, they represent a completely different level of information about the speaker and hence can be considered independent in terms of the information content.

Usually, static and delta cepstra are concatenated to represent a single feature vector for the task of speaker recognition. This is referred to as fusion at the feature level [5]. It is, however, reported in the literature that the fusion strategies work best at the score level [2]. Hence in this study, the fusion of the information obtained from static and delta cepstra is considered at the score level.

Various score level fusion schemes are evaluated in this study. Amongst these, the Support Vector Machine (SVM) is of particular interest. The use of Support Vector Machines in speaker verification has been considered relatively recently. To date, however, SVM have only been implemented at the feature level for speaker verification [6]. In this approach, the feature space is projected into some different hyperspaces so that the discrimination between the true and impostor speaker utterances is maximised. It has also been shown that combining SVM and GMM would lead to improvement in discrimination capability. [6]. In the present work, SVM are used at the score level (to combine the likelihood scores obtained from the static and delta cepstra) with the aim to maximise the separation of the true and impostor speakers. The rest of the paper is structured as follows. Section 2 gives the theory of various fusion schemes. Section 3 details the experimental setup. Section 4 discusses the results, whilst Section 5 presents the overall conclusions

## 2. FUSION TECHNIQUES

### 2.1. Weighted Average Fusion

In weighted average schemes, the fused score for each class (e.g. *j*) is computed as a weighted combination of the scores obtained from *N* matching streams as follows.

$$f_j = \sum_{i=1}^{N} w_i x_{ij} \quad , \qquad (1)$$

where, $f_j$ is the fused scores for $j^{th}$ class, $x_{ij}$ is the normalised match score from the $i^{th}$ matcher and $w_i$ is the corresponding weight in the interval of 0 to 1, with the condition

$$\sum_{i=1}^{N} w_i = 1 \,, \qquad (2)$$

There are three sub-classes of this scheme, which primarily differ in the method used for the estimation of weight values.

*2.1.1. Brute Force Search (BFS)*

This approach is based on using the following equation [5].

$$f_j = x_j^1 * a + x_j^2 * (1-a) \qquad , \qquad (3)$$

where $f_j$ is the $j^{th}$ fused score, $x_j^p$ is the $j^{th}$ normalized score of the $p^{\text{th}}$ matcher, $p = 1,2$ and $0 \le a \le 1$.

*2.1.2. Matcher Weighting using FAR and FRR (MW – FAR/FRR)*

In this technique the performance of the individual matchers determines the weights so that smaller error rates result in larger weights. The performance of the system is measured by False Acceptance Rate (FAR) and False Rejection Rate (FRR). These two types of errors would be computed at different thresholds. Threshold that minimises the absolute difference between FAR and FRR on the development set is then taken into consideration. The weights for the respective matchers are computed as follows [7].

$$w_u = \frac{1-(FAR_u + FRR_u)}{2-(FAR_v + FRR_v + FAR_u + FRR_u)} \,, \qquad (4)$$

where $u$=1, 2, $v$ =1, 2 and $u$ is not equal to $v$ with the constraint $w_u + w_v = 1$

The fused score using different matchers is given as

$$f_j = w_u * x_j^u + w_v * x_j^v \qquad (5)$$

where, $w_k$ is the weight from the $k^{\text{th}}$ matcher, $x_j^p$ is the jth normalised score of matcher p and $f_j$ is the fused score.

*2.1.3. Matcher Weighting based on EER (MW - EER)*

The matcher weights in this case depend on the Equal Error Rates (EER) of the intended matchers for fusion. EER of matcher $m$ is represented as $E^m$, $m$=1, 2 and the weight $w_m$ associated with matcher $m$ is computed as [8].

$$w_m = \frac{1}{E^m \left( \sum_{m=1}^{M} \frac{1}{E^m} \right)} \qquad (6)$$

Note that $0 \le w_m \le 1$, with the constraint given in (2). It is apparent that the weights are inversely proportional to the corresponding errors in the individual matchers. The weights for less accurate matchers are lower than those of more accurate matchers. The fused score is calculated in the same way as in equation (1).

## 2.2. Fisher Linear Discriminant (FLD)

In FLD, the linear boundary between the data from two classes is obtained by projecting the data onto the one dimensional space [9].

For data $x$, the equation of the boundary can be given as

$$h(x) = w^T x + b \,, \qquad (7)$$

where, $w$ is a transformation matrix obtained on the development data and $b$ is a threshold determined on the development data to give the minimum error of classification in respective classes. The rule for class allocation of any data vector is given by

$$x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \text{ if } w^{\mathrm{T}} x + b \begin{cases} > \\ < \end{cases} 0 \qquad , \qquad (8)$$

*2.2.1. Training the FLD*

Given a range normalised data $x_i$ from class Ci having a multivariate Gaussian distribution with the statistics $[m_i, S_i], i \in 1 and 2$, where $S_i$ and $m_i$ are a scatter matrix and mean for the particular class i. The scatter matrix is given as [9]

$$S_i = \sum_{k \in C_i} (x_k - m_i)(x_k - m_i)^T \,, \qquad (9)$$

where, $T$ is a transpose operation.

The overall within class scatter matrix $S_W$ and the between class scatter matrix $S_B$ are given by

$$S_W = \sum_{i=1}^{2} S_i \qquad (10)$$

$$S_B = (m_2 - m_1)(m_2 - m_1)^T \qquad (11)$$

The transformation matrix $w$ is obtained using the equation

$$w = S_W^{-1} (m_2 - m_1) \qquad (12)$$

## 2.3. Quadratic Discriminant Analysis (QDA)

This technique is the same as FLD but is based on forming a boundary between two classes using a quadratic equation given as [10]

$$h(x) = x' Ax + b' x + c \qquad (13)$$

For training data 1 and 2 from two different classes, which are distributed as $N[m_i, \Sigma_i], i \in 1 and 2$, the transformation parameters A and b can be obtained as

$$A = -\frac{1}{2}\left(\Sigma_1^{-1} - \Sigma_2^{-1}\right) \qquad (14)$$

$$b = \Sigma_1^{-1} m_1 - \Sigma_2^{-1} m_2 \qquad (15)$$

The classification rule in QDA is of the same nature as in FLD, only the equation is replaced appropriately.

## 2.4 Logistic Regression (LR)

The assumption in this technique is that the difference between log likelihood functions from two classes in data $x$ is linear in $x$ [9].

$$\log\left(\frac{p(x/\omega_1)}{p(x/\omega_2)}\right) = \alpha + \beta^T x \qquad (16)$$

Parameters in the above equation can be calculated with the maximum likelihood approach with an iterative optimisation scheme on some development data. Details can be found in [9].

The allocation rule for the test data is given as

$$x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \; if \;\; \alpha_0 + \beta^T x \begin{cases} > \\ < \end{cases} 0 \qquad (17)$$

## 2.5 Support Vector Machines (SVM)

SVM is a classification technique based on forming a hyper plane that separates data from two classes with a maximum possible margin. SVM is based on the principle of Structural Risk Minimization (SRM) [11]. SRM principle states that better generalization capabilities are achieved through a minimization of the bound on the generalization error. The SVM uses the following function to map a given vector to its label space (i.e., -1 or +1)

$$f(x) = \text{sign}\left(\sum_{i=1}^{l} a_i y_i k(x, x_i) + b\right) \qquad (18)$$

where $k(x, x_i)$ is a kernel function that defines the nature of the decision surface that separates the data, $x$ is the input vector of a test set, $x_i$ is the input vector of the i[th] training example, $l$ is the number of training examples, $b$ is a bias estimated on the training set, $y_i$ is the class specific mapping label and $a_i$ are the solutions of the following Lagrangian in the quadratic programming problem.

$$Q(a) = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j y_i y_j k(x_i, x_j) \qquad (19)$$

with the constraints,

$$\sum_{i=1}^{l} a_i y_i = 0 \qquad (20)$$

More details of this equation are given in [11]. In the resulting solution, most $a_i$ are equal to zero, which refer to the training data that are not on the margin. The training examples with non-zero $a_i$ are called support vectors, which are the input vectors that lie on the edge of the margin. Introducing new data outside of the margin will not change the hyper plane as long as the new data are not in the margin or misclassified. Therefore, the classifier must remember those vectors which define the hyper plane.

The kernel function $k(x, x_i)$ can have different forms. More details can be found in [11]. In this work, linear and polynomial kernel functions with a degree of 2 (quadratic) are used. These are given by following equations,

$$Linear: k(x, x_i) = x^T x_i \quad , \qquad (21)$$

$$Quadratic: k(x, x_i) = (x^T x_i + 1)^2 , \qquad (22)$$

## 2.6. Range-Normalisation Techniques

Range-normalisation is the task of bringing raw scores from different matchers to the same range. This is a necessary step in any fusion system as fusing the scores without such normalisation would de-emphasise the contribution of the matcher having a lower range of scores. Two different normalisation techniques have been evaluated in this paper [8].

### *2.6.1 Min-Max Normalisation (MM)*

This method uses the following equation

$$x = \frac{n - \min(n)}{\max(n) - \min(n)} \quad , \qquad (23)$$

where, $x$ is the normalised score, $n$ is the raw score, and max and min functions specify the maximum and minimum end points of the score range respectively.

### *2.6.2. Z-score Normalisation (ZS)*

This method transforms the scores having some Gaussian distribution to a standard Gaussian distributional form. It is given as

$$x = \frac{n - mean(n)}{std(n)} , \qquad (24)$$

Where, $n$ is any raw score, and mean and std are the statistical mean and standard deviation operations.

# 3. EXPERIMENTAL SETUP

## 3.1. Speech Data

The speech data used in this work is from the TIMIT database. Material from all the 630 speakers is used. For each speaker, the utterances 'sa1' and 'sa2' are used for the development and testing respectively. The rest of the 8 utterances for each speaker are used for developing the speaker representation as a Gaussian Mixture Model (GMM) with 32 components.

## 3.2. Feature Extraction

The extraction of cepstral parameters is based on first pre-emphasising the input speech data using a first order digital filter with a coefficient of 0.95 and then segmenting it into 20 ms frames at intervals of 10 ms using a Hamming window. 16 LPCC coefficients are then obtained via a linear prediction analysis. For obtaining MFCC, speech spectrum for each frame is weighted by a Mel scale filter bank. The discrete cosine transformation of the log magnitude outputs of these filters gives the MFCC for that speech frame. For each type of cepstra, a polynomial fit method is used to obtain the delta coefficients [4].

## 3.3. Testing

The scores generated with the development utterances are first used to obtain the training parameters in various fusion techniques. True and impostor scores from static and delta streams are pooled and then normalised according to the chosen range-normalisation scheme. Parameters obtained in the fusion schemes are then used in the test phase to transform the normalised test scores according to the fusion scheme. The

verification performance is then obtained on the transformed scores in terms of equal error rates (EER) via the DET curves

.

## 4. RESULTS AND DISCUSSIONS

The experimental results are presented in the following tables. It can be seen that (in most cases) the ZS normalisation is exhibiting more effectiveness than the MM normalisation. In some cases though, the two approaches provide comparable performance.

It can be observed that the way fusion techniques work for combining the static and delta features is not identical in the two considered cases of LPCC and MFCC. In the case of LPCC features, improvements are seen in majority of the fusion cases by fusing the scores from static and delta features as compared to the feature level concatenation. In some cases such as Linear SVM and LR, the results are even better than using the individual feature streams. Thus under this experimental setup, LR and Linear SVM give the best results for LPCC features. In the case of MFCC data, all of the fusion techniques except MW-EER indicate that score level fusion can give better performance than the feature level concatenation. But no fusion techniques for MFCC are seen to exceed the performance of the baseline MFCC static features. Thus the best results obtained in this case are still with MFCC static features.

Thus it can be said that the speaker verification systems can benefit through the score level fusion, but this depends on the types of feature as well as the normalisation method used.

| Cepstra | EER % | Cepstra | EER % |
|---|---|---|---|
| LPCC static (s) | 1.76 | MFCC static (s) | 2.06 |
| LPCC delta (d) | 39.64 | MFCC delta (d) | 38.89 |
| LPCC (s + d) | 2.44 | MFCC (s + d) | 3.23 |

**Table 1**: Baseline Results

| EER % | BFS | MW (FAR/FRR) | MW - EER | FLD |
|---|---|---|---|---|
| LPCC (s + d) | 2.17 | 3.17 | 2.16 | 10.64 |
| MFCC (s + d) | 2.21 | 4.45 | 2.45 | 2.27 |

| EER % | QDA | LR | SVM Linear | SVM Poly |
|---|---|---|---|---|
| LPCC (s + d) | 4.60 | 3.87 | 4.20 | 4.39 |
| MFCC (s + d) | 2.27 | 2.73 | 2.32 | 3.17 |

**Table 2**: Score Level Fusion (MM Normalisation)

| EER % | BFS | MW (FAR/FRR) | MW - EER | FLD |
|---|---|---|---|---|
| LPCC (s + d) | 1.74 | 2.70 | 2.06 | 2.85 |
| MFCC (s + d) | 2.22 | 3.83 | 2.30 | 2.27 |

| EER % | QDA | LR | SVM Linear | SVM Poly |
|---|---|---|---|---|
| LPCC (s + d) | 1.75 | 1.14 | 1.08 | 1.71 |
| MFCC (s + d) | 2.27 | 2.79 | 2.34 | 2.65 |

**Table 3**: Score Level Fusion (ZS Normalisation)

## 5. CONCLUSIONS

It can be concluded from this study that the combination of complementary information from the speech static and delta cepstra can improve the performance in the speaker verification. Improvements are of greater extent in the case of LPCC features. In this case, the fusion of the information at the score level is more effective than that at the feature level. Amongst various fusion methods considered, SVM approach has appeared to provide the best performance in terms of reducing error rates in speaker verification. Finally the ZS normalisation method exhibits better performance than MM normalisation for the fusion task.

## 6. REFERENCES


1. Fabio Roli et.al. “An Experimental Comparison of Classifier Fusion Rules for Multimodal Personal Identity Verification Systems”, Proc. Multiple Classifier Systems, Springer-Verlag, 2002, pp. 325-336.

2. Conrad Sanderson and Kuldip K. Paliwal, “Identity Verification using Speech and Face Information”, Digital Signal Processing, 2004.

3. D. Reynolds and R. Rose, “Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models”, IEEE Transactions on Speech and Audio Processing, 3 (1), Jan. 1995.

4. D. O'Shaughnessy, Speech Communication: Human and Machine, Addison -Wesley, 1987

5. Ariyaeeinia A. M. and Sivakumaran P., “Effectiveness of Orthogonal Instantaneous and Transitional Feature Parameters for Speaker Verification”, 29th Annual Carnahan Conference on Security Technology, 1995, pp. 79 – 84.

6. V. Wan and S. Renals, IEEE International Workshop on Neural Networks for Signal Processing 17 - 19 September 2003.

7. Y. Wang, T. Tan and A. K. Jain, “Combining Face and Iris Biometrics for Identity Verification”, Proceedings of Fourth International Conference on AVBPA, (Guildford, U. K.), pp. 805-813, June 2003

8. M. Indovina et.al., “Multimodal Biometric Authentication Methods: A COTS Approach”, Proc. MMUA 2003, Workshop on Multimodal User Authentication, Santa Barbara, CA, December 11-12, 2003.,pp. 99-106

9. C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, New York, 1996.

10. B. Flury, Common Principle Components and Related Multivariate Models, John Wiley and Sons, USA, 1988.

11. C.J.C. Burges, “A tutorial on support vector machines for pattern recognition”. Data Mining and Knowledge Discovery, 2(2),pp. 955-974, 1998

# THE BIOSEC MOBILE USABILITY LAB FOR BIOMETRIC SECURITY SYSTEMS

*Felix Eschenburg, Gary Bente, Heide Troitzsch, Rafal Powierski & Oliver Fischer*

Department of Social and Differential Psychology, University of Cologne
felix.eschenburg@uni-koeln.de

## ABSTRACT

Making biometrics an integral part of internet applications could significantly reduce security problems that arise from incorrect user behavior, but it does not eliminate the human factor entirely. User acceptance and the system's usability will play a crucial role for the technologies' success. This paper presents the mobile usability lab developed within workpackage 6 of the integrated project BioSec (see http://www.biosec.org). The mobile lab allows us to gather a combination of objective and subjective data regarding user acceptance and usability. Data collection can take place both online and offline. Results indicate that the chosen multimodal research approach is useful in identifying specific usability problems .

## 1. INTRODUCTION

Nowadays the internet is not only used to find, gather and share information, but also to give users remote access to bank accounts, online shops and secured networks. Users' identity is usually verified my means of passwords and PINs . These methods can be a serious threat to the system's security, because users often find it difficult to memorize different passwords and are generally not particularly motivated to behave in a manner that ensures system security [1]. Biometrics are seen as a potential solution to this problem and are expected to provide a higher level of security for these remote applications in the near future. These expectations resulted in a remarkable increase in R&D efforts [2].

Nevertheless, the implementation of biometric security systems will not eliminate the human factor. Understanding the psychological processes that underlie the users' behavior will play a major role for the technology's success [3]. This refers not only to the way the user interacts with the system, but also to the way the system is perceived and to the attitudes the user has towards the system. In essence, we predict that user acceptance and system usability will be crucial for the technology's success.

Several approaches have been adopted to investigate user acceptance and usability. A large proportion of research has focused on ATM applications [4,5]. Results indicated that user acceptance increases after the system had been used for a while, although some problems concerning the system's usability were identified. A generally high acceptance to implement biometrics in e-banking and e-transaction applications was also found by Giesing [6] among South African participants, whereas Rodriguez et al. [7] presented a first approach to overcome usability problems by using an interface based on an embodied conversational agent.

Although several studies have tried to use subjective as well as objective data [4,5,7], no integrated research concept has been presented yet. In this paper we will present the BioSec mobile usability lab as a multi-modal research platform that will be applicable to a wide variety of settings (field and laboratory), sensor technologies, applications, and research questions (general acceptance or specific questions regarding concrete handling of particular sensor techniques).

## 2. THE BIOSEC MULTIMODAL SETUP

The BioSec mobile lab consists of a specific selection of methods to obtain subjective and objective data. The mian advantage of such a multimodal approach is that the specific drawbacks of subjective and objective measurement methods can be compensated to a large extent, which in turn results in more reliable data.

### 2.1. Subjective measurement methods

Questionnaires are an economic way of gathering subjective data, with non-standardised questionnaires being particularly common in this field of research [8]. The reason for the use of non-standardised questionnaires is that they can be tailored to the tested system. However, the low level of standardization means that research criteria are very difficult to establish and results are difficult to compare across studies [9].

The BioSec acceptance and usability questionnaire consists of three parts. The pre-questionnaire is meant to be given before concrete experiences with biometrics have been made by the user. It aims at different aspects, such as concerns for health and hygiene, general attitudes and acceptance, perceived reliability, prior knowledge and privacy concerns. A short questionnaire assessing the subject's attitudes regarding the enrolment procedure has also been developed and is handed out directly after the enrolment procedure. The post-questionnaire assesses user satisfaction, learnability and ease of use after the subjects have used a biometric security system.

Items were generated from checklists that were in turn based on an extensive literature review. The selected items were validated by an expert panel to ensure construct and face validity. Each part of the questionnaire contained a

combination of open questions and items that used 5-point Likert scales.

## 2.2. Objective data

The main advantage of objective data is that they are not prone to biases such as social desirability or memory. Our approach integrates psycho-physiological, behavioural and eye movement data. This required the development of a specific platform, which is described below, followed by a detailed account of each objective measurement.

*Integration of internet applications, biometric sensors and authentication software*

The first step was to develop an application that scripts the experimental process and triggers the different biometric devices such as fingerprint-sensor or iris-camera. This software is needed to use the different authentication-devices together with the provided software and user interfaces in a controlled experimental environment.. The tool is able to start any authentication-device by integrating a provided SDK, or by simulating an event needed to start the intended application (for example a hidden application simulating a required password input). It also makes it possible to present users any message before and/or after the authentication-process, or to repeat the process if required. The latter is particularly important, as it permits to experimentally vary the error rates of the system.

*Collecting behavioral data with the Mediascore browser*

Based on the software described above a browser has been developed by Mediascore, further referred to as the Mediascore browser. It allows us to capture the screen content - including mouse movements - during the whole login process. A script that controls a common web cam to record the user's behavior is also embedded into the browser. The so gathered behavioral data always refers to a fixed time frame.

*Psycho-physiological data*

Within the mobile lab, psycho-physiological measures can be collected by means of a multi-channel mobile device. This innovative technology called SMARD-Watch® was developed by the Department of Stress Research, Berlin, and provides a non-invasive and simultaneous measurement of various psycho-physiological responses, ranging from heart rate, pulse rate and electro-dermal activity to skin-conductance and electromyogram. The necessary sensors are integrated into a small strap that resembles a watch and can be worn around wrist or ankle.

The SMARD-Watch's greatest advantage is its non-invasiveness. A clear disadvantage, however, is the SMARD-Watch's low resolution in time. The device does not permit to elicit data on a second or even millisecond basis, but only makes one measurement approximately every ten seconds.

*The gaze tracking system*

Given that gaze direction is an indicator of the users' focus of attention, eye tracking can help identify the causes of specific orientation and handling problems [11,12]. The measurement of time spent at different areas of the screen, for example, can be instrumental in understanding which information is accessed quickly and which not, as well as where the user searches for further information [12].

Our Mediascore gaze tracking system consists of two cameras, four infrared LEDs and a mirror. The cameras are fixed on a headband. One camera is a conventional video camera and is aligned in a way such that it records the user's field of view. The other camera is applied between the four infrared LEDs and is directed at the infrared mirror. The mirror is fixed at an angle of approximately 45° in front of the user's eye. With this setup, the infrared camera captures the position of the eye at a rate of 25 Hz. The LEDs are used as a reference signal. This signal is necessary to eliminate errors that are due to minimal displacements of the headset. If the LEDs were not used, even marginal headset displacements would result in a large display error of the eye position.

## 2.3. The BioSec laboratory study

The laboratory study was designed within BioSec to conduct a first set of usability tests of existing biometric security devices. It was also used to test the mobile usability lab.

The design aimed at comparing two different biometric modalities and two different sensors that use the same biometric modality. 63 subjects matched for age (half the subjects were between 20 and 30 years, the other half between 40 and 50 years of age) and gender were invited to the Mediascore laboratories to test one of three biometric systems (see table 1 for the design).

| | **Task: Online Banking** | | |
|---|---|---|---|
| **Tech-nology** | **Fingerprint** ***BiometriKa*** | **Fingerprint** ***Siemens*** | **Iris scan** ***Pana-sonic*** |
| **Sub-jects** | 22 | 22 | 23 |
| |  |  |  |

**Table 1:** Design of BioSec first phase usability testing.

All subjects were randomly assigned to the technologies, but were matched for age and gender within experimental conditions. Participants were asked to fill-in the pre-questionnaire before enrolling with one of the three systems. After that they had to a) fill-in the enrolment questionnaire, b) login to an online-banking application and c) fulfil a given task (transfer some money from one account to another). At the end, participants were given the post-questionnaire and answered a few semi-structured interview questions.

Pre-tests showed that the Smard-Watch's low time resolution made data interpretation very difficult. We therefore decided not to use this device in the main study. This issue will be further discussed in the fourth section of this paper. A setup using the complete mobile lab is presented in figure 1.

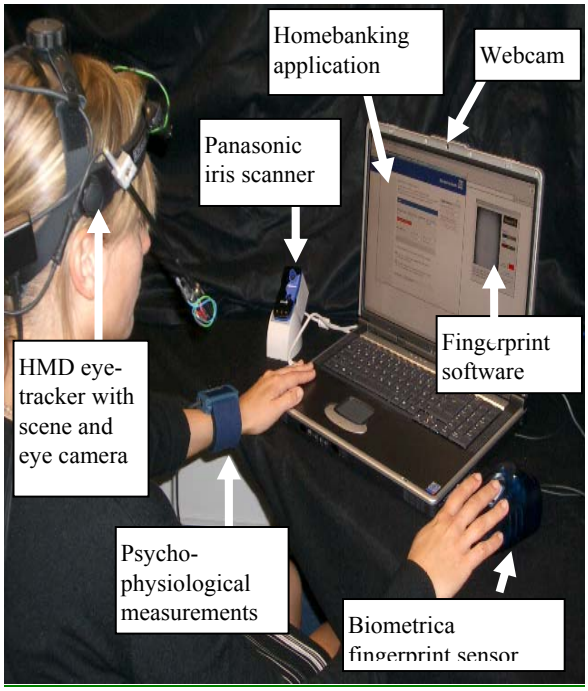



**Figure 1**: The BioSec mobile lab

## 3. EXPERIMENTAL RESULTS

Separate factor analyses were conducted for pre- and post questionnaires. The factorial structure of these two questionnaires differed substantially, indicating that the dimensions of participants' technology perceptions changed as a result of technology use. Internal consistency (Cronbach's Alpha) for the scales we had developed during the early stages of the project was also low. This, however, made a systematic dimensional interpretation difficult to justify demanded that all our analyses had to be conducted on a single item basis.

The items pertaining to prior knowledge also turned out to be insufficient, because they produced no variance (nearly all participants indicated that they had heard of biometric systems, but none had seen or used one).

### 3.1. Acceptance of biometric security is high

Our results are in line with those found in earlier studies [4,5]. General acceptance of biometric security systems is high and the usage of the systems produces a shift of opinion towards an even higher acceptance. About two thirds of the participants answered the question whether their "*attitude towards biometric systems changed since using them*" with either "*mostly positive*" or "*positive*". A one-sample T-test against the scale's mean value of 3 was highly significant ($T = 9{,}578$; $p < 0{,}001$).

Pre-post comparisons also showed that the users expressed a more positive attitude towards biometrics after they used the systems. They perceived them as easier to use, more reliable, less inconvenient and faster to use (we computed ANOVAs using a 2x2x3x2 fractional factorial design reduced to degree 2; all $F > 6.45$, all $p < 0.05$). See figure 2 for details.

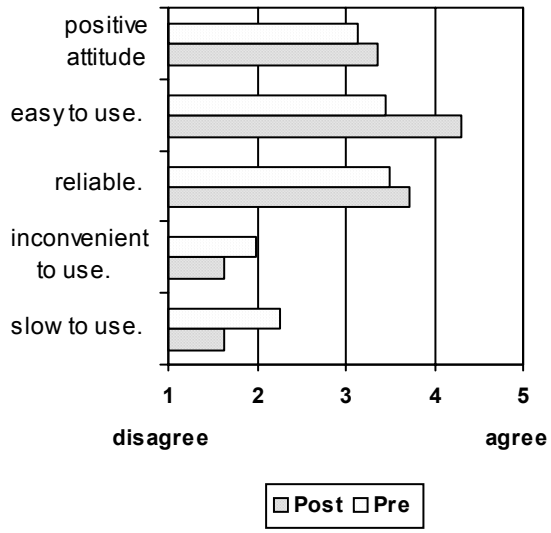



**Figure 2**: Some items regarding user acceptance: Pre-Post comparison

### 3.2. Usability problems with Iris scanner

The subjective data have been analysed by computing an ANCOVA with the time the users needed to fulfil the (dummy-) task as a covariate.

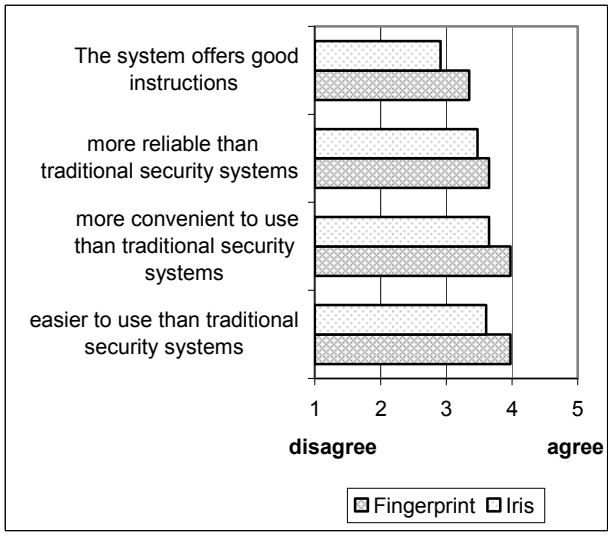



**Figure 3**: Some items regarding usability

The iris recognition system was perceived as more difficult to use, less convenient to use and less reliable than the fingerprint sensors (see figure 3). Also, the Panasonic interface was perceived as less helpful. All effects were significant on the 5% level (all $F>4.39$, all $p<0.05$). No significant differences between the two fingerprint devices were found.

To obtain deeper insights into usability problems first analyses of the gaze trecking data with a reduced set of subjects were computed. The gaze pattern points to severe problems of focusing attention on the iris scanner when there is a task pending on the computer screen. People seem to be impatient to continue the login process and they expect a positive

feedback via the computer screen. The sequential attention shift that is required here seems to be most irritating for the subjects (see figure 4).

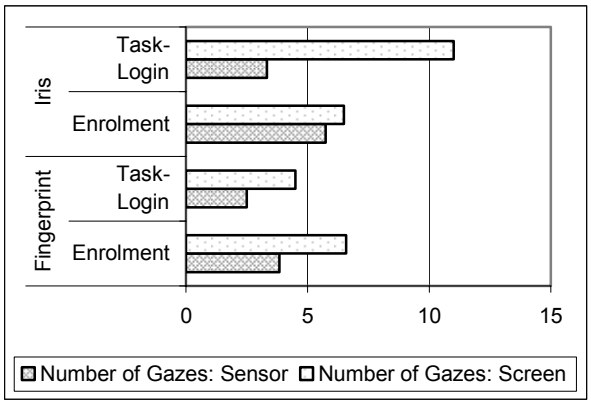



**Figure 4**: Number of gazes fingerprint vs. iris

# 4. CONCLUSIONS

Overall, the chosen multimodal research approach proved useful in identifying usability problems and different determinants of user acceptance. A pre-post comparison of items regarding general attitudes showed that most users were positively surprised how easy the devices were to use. Actually using a biometric system instead of only hearing about it can obviously help to reduce prejudices against the technology. Additionally, the gaze tracking data proved to be useful to identify severe usability problems.

However, our study also revealed some serious research deficits. Firstly, the chosen device to gather psycho-physiological data (i.e., the SMARD-Watch) could not be used to detect the kind of short arousals that are typically related to emotional responses during the login process. Unfortunately, other methods to gather different kinds of psycho-physiological data are still invasive or they do not offer the necessary time resolution. The development of a device similar to the SMARD Watch, but capable of measuring at a higher resolution of time, would be a great benefit to media research.

Secondly, the instability of the internal structure of our questionnaire highlights the need for a standardized research questionnaire to assess user acceptance and usability of biometric systems. Without such an instrument, comparability across studies, and seemingly even between conditions will be extremely limited. Additionally, a standardized questionnaire consisting of several scales that aim at measuring particular psychological constructs would permit to develop more sophisticated theories about the determinants of user acceptance and their inter-dependencies. Our future research will aim at the development of such a standardized questionnaire.

# 5. REFERENCES


1.Adams, A. & Sasse, M. A. (1999) Users are not the enemy. Communications of the Art. 42 (12), 41-46. available at: <portal.acm.org/citation.cfm?id=322806> [Accessed 1th June, 2005]

2.Petermann, T. & Sauter, A. (2002). Biometrische Identifikationssyysteme – Sachstandsbericht. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag, Arbeitsbericht Nr.76, Februar 2002.

3.Ashbourn, J. (2004). User Psychology. In: Ashbourn, J. Practical Biometrics. From Aspiration to Implementation. Heidelberg: Springer

4.Coventry, L., De Angeli, A., and Johnson, G. (2003) Usability and Biometric Verification at the ATM Interface. Proceedings of HCI 2003, 153-160. ACM Press, New York, NY, USA. Available at: < http://portal.acm.org/citation.cfm?id=642639 > [Accessed 15th November, 2004]

5.Behrens, M. & Roth, R. (2002). BioTrusT: Untersuchung der Akzeptanz und Nutzung biometrischer Identifikationsverfahren. In Nolde, V. & Leger, L. (Eds.) Biometrische Verfahren – Körpermerkmale als Passwort – Grundlagen, Sicherheit, und Einsatzgebiete biometrischer Identifikation. Köln: Fachverlag Deutscher Wirtschaftsdienst.

6.Giesing, I. (2003). User perception related to identification through biometrics within electronic business. Master's Dissertation, available at: http://upetd.up.ac.za/thesis/available/etd-01092004-141637/ [Accessed 15th November, 2004]

7.Rodríguez, R.G., Trapote, A.H., Pozo, R. F., de la Morena Sanz, A. and Gómez, L. H. (2004). Distributed Environment for Multimodal Biometric Evaluation in Internet Applications. 2nd COST 275 Workshop – Biometric on the Internet, Vigo, 25-26 March 2004

8.Faulkner, C. (2000). Usability Engineering. UK: Macmillian Press LTD.

9.Bortz, J. & Döring, N. (1995): Forschungsmethoden und Evaluation für Sozialwissenschaftler. Berlin: Springer.

10. Kempter, G. & G. Bente, (2004). Psychophysiologische Wirkungsforschung: Grundlagen und Anwendungen. In: G. Bente, R. Mangold, R. Vorderer (Hrsg). Lehrbuch der Medienpsychologie (272-295).Göttingen: Hogrefe

11. Aaltonen, A. (1999). Eye tracking in usability testing: Is it worthwhile? Presented at Workshop in CHI´99 on Usability & Eye Tracking: ACM Conference on Human Factors in Computing Systems, Pittsburgh, PA. Online-Document (Accessed 25th March 2003)

12. Bente, G. (2004). Erfassung und Analyse des Blickverhaltens. In: G. Bente, R. Mangold, R. Vorderer (Hrsg). Lehrbuch der Medienpsychologie. Göttingen: Hogrefe

# ON EMERGING BIOMETRIC TECHNOLOGIES

*Georgios Goudelis, Anastasios Tefas and Ioannis Pitas*

Department of Informatics Aristotle University of Thessaloniki

pitas@aiia.csd.auth.gr

## ABSTRACT

Many body parts, personal characteristics and imaging methods have been suggested and used for biometrics systems: fingers, hands, feet, faces, eyes, ears, teeth, veins, voices, signatures, typing styles and gaits. Literature contains a wide range of techniques occupying a large number of implemented algorithms regarding biometrics. In this paper we will introduce the latest improvements on biometric systems. A distinct dissociation separates them to intrusive and non-intrusive according to the level of nuisance that each system sets off.

## 1. NON INTRUSIVE SYSTEMS

The amount of non-intrusive systems so far is not so large. Most of the developed techniques require the imminent participation of the person that is to be recognized. Although that voluntary presence seems to agree more with the idea of protection of the personal data, intrusive methods are not always the requisite ones.

The latest achievements on non-intrusive biometrics present knew technologies that promise to change the way of thinking in this direction. Thermogram, smile identification, lip recognition and hyperspectral analysis seem to be the most important and promising techniques.

### 1.1 Thermogram

Scientists have found that a unique heat distribution pattern can be obtained in human face. This pattern can be seen by taking pictures using infrared cameras. The different densities of bone, skin, fat and blood vessels all contribute to an individual's personal "heat signature". Conventional video cameras sense reflect light so that image values are a product of both intrinsic skin reflectivity and external incident illumination, thus obfuscating the intrinsic reflectivity of skin. Thermal emission from skin, on the other hand, is an intrinsic measurement that can be isolated from external illumination.

Nine different comparative parameters are used excluding the nose and ears, which are prone to wide variations in temperature. Once a picture of a face is taken, its thermal image can be matched with accuracy against a database of pre-recorded thermographs.

This work is funded by the integrated project BioSec IST-2002-001766 (Biometric Security, http://www.biosec.org), under Information Society Technologies (IST) priority of the 6th Framework Programme of the European Community.

A study in [1] examines the invariance of Long-Wave Infrared (LWIR) imagery with respect to different illumination conditions from the viewpoint of performance comparisons of two well known face recognition algorithms applied to LWIR and visible imagery. A rigorous data collection protocol is developed that formalize face recognition analysis for computer vision in the thermal IR.

One of the obvious advantages of thermal imagery is the ability to operate in complete darkness which makes it ideal for covert surveillance. However, it has other limitations including that it is opaque to glass [2].

### 1.2 Smile Recognition

Another promising method for person recognition is suggested in [3]. A high speed camera with a strong zoom lens allows smile maps to be produced. This map is claimed to be unique for each person. This new method compares images of a person, taken fractions of a second apart, while they are in the smiling process. The system probes the characteristic pattern of muscles beneath the skin of the face. The way the skin around the mouth is moved between shots is analysed by tracking the change position and direction of tiny wrinkles in the skin. The data is used to produce an image of the face overlaid with tiny arrows that indicate how different areas of skin move during a smile. This movement is controlled by the pattern of muscles under the skin and is not affected by the size of the smile or the presence of make-up.

The system has been successfully tested so far on a very small database consisted of 4 lab members smiling samples. The system is currently tested on a larger group of 30 smiling faces but it is obviously too early to evaluate its robustness.

### 1.3 Lip Recognition

Another study concerning human recognition is described in [4]. A lip recognition method that uses shape similarity when vowels are uttered is proposed. In this method, a mathematical morphology analysis is applied using three different structuring elements. The proposed structuring elements are the square and vertical and horizontal line and they are used for deriving a pattern spectrum. The shape vector is compared with the reference vector to recognize an individual from its lip shape.

Experimental results show that the shape vector contains information capable to perform recognition by lip shape. In particular eight Japanese lips could be classified with 100.0% accuracy.

Of course the result is fictitious and authors make that clear. They note that the system is not sophisticated yet and

classification accuracy has to be improved by considering a new structuring element (for instance, rectangle, ellipse or asymmetric shape). Collection of a significantly larger database is incontestably required.

## 1.4 Hyperspectral Images

Hyperspectral cameras provide useful discriminates for human face recognition that cannot be obtained by other imaging methods [5]. The use of near-infrared hyperspectral images for the recognition of faces is examined, over a database of 200 subjects. The hyperspectral images were collected using a CCD camera equipped with a liquid crystal tunable filter to provide 31 bands over the near-infrared (0.7_m-1.0_m).

It is experimentally demonstrated that this algorithm used in [5], is able to recognize faces over time in the presence of changes in facial pose and expression. The authors claim, that the algorithm performs significantly better than current face recognition systems for identifying rotated faces. Performance might be further improved by modeling spectral reflectance changes due to face orientation changes.

# 2. INTRUSIVE SYSTEMS

The majority of developed biometric systems belong in the category of intrusive systems. These systems require the cooperation of the subject to be recognized. The level of intrusiveness is determined by the level of cooperation that they demand from the subject. Many well-known biometric modalities like iris, retina, fingerprints, keystroke dynamics and dental analysis are representative examples. In the following the latest developments within this class of biometric recognition/verification techniques are presented.

## 2.1 Electrocardiogram (ECG)

An electrocardiogram (ECG / EKG) is an electrical recording of the heart and is used in the investigation of heart diseases. Recently, several researchers characterized the ECG as unique to every individual [6], [7], [8].

In [9] the ECG processing followed a logical series of experiments with quantifiable metrics. Data filters were designed based upon the observed noise sources. Fiducial points were identified on the filtered data and extracted digitally for each heartbeat. From the fiducial points, stable features were computed that characterize the uniqueness of an individual. The tests show that the extracted features are independent of sensor location, invariant to the individual's state of anxiety, and unique to an individual.

The dataset was used to identify a population of individuals. Additional data collection is being tried in order to test the scalability of the features to characterize a large population as well as the stability of those features over long time intervals.

## 2.2 Mouse Dynamics

It is known that most of the current available technologies typically require specialist, and often expensive equipments that hindering their widespread and distribution. An advantageous on this matter solution comes to give the University of Queen Mary in London [10]. The new technique is based on mouse dynamics.

According to the researchers of Queen Mary University the software produced, uses state of the art pattern recognition algorithms combined with artificial intelligence to provide a biometric layer over traditional password based security. The system learns an optimum set of mouse-movement characteristics unique to the user's mouse-written signature and uses them to authenticate later signatures. The artificial intelligence can also learn over time to include changes of the user's typing and mouse signature characteristics.

The specific method is mostly suggested as an on-line biometric verification solution. On-line banking, shopping, or accessing web based e-mail could be a few of its possible application. In addition the technique can be used to validate computer-controlled entry to rooms or buildings, confirming identity at security checkpoints and so on without expensive specialist equipment.

## 2.2 Finger-Vein Patterns

Another method for personal identification is proposed in [11] and is based on finger-vein patterns. The authors proposed a scheme based on finger vein patterns as a scheme of biometric identification utilizing biological information. Since the finger vein images taken to obtain finger vein patterns are obtained by irradiating the fingers with infrared rays, fluctuations in brightness due to variations in the light power or the thickness of the finger occur. The proposed scheme is robust against brightness fluctuations, compared with the conventional feature extraction schemes. The evaluation results of the proposed scheme when applied to a person identification system showed an effective error rate of 0.145%.

Researchers in [11] argue that an image of a finger captured under infrared light contains not only the vein pattern but also irregular shading produced by the various thicknesses of the finger bones and muscles. The proposed method extracts the finger-vein pattern from the unclear image by using line tracking that starts from various positions. The extraction of the patterns of an unclear original image, line-tracking operations with randomly varied start points are repeatedly carried out.

It is reported too, that the mismatch ratio is slightly higher during cold weather because the veins of the finger can become unclear. Therefore, a devise that can capture the vein pattern more clearly and a feature extraction algorithm that is robust against these fluctuations should be investigated.

It is worth noting that "Hitachi engineering Co. Ltd", has already published the commercial product called "SecuaVeinAttestor". Full specification and characteristics of this product are given in [12].

## 2.3 Ear Prints

Using ears in identifying people has been subject of investigation for at least 100 years. The researches still discuss if ears are unique or unique enough to be used as biometrics. Ear shape applications are not commonly used yet, but the

area is interesting especially in crime investigation. Burge and Burger think that ear biometrics is a “viable and promising new passive approach to automated human identification” [13].

Ear data can be received from photographs, video or earprints produced by pressing the ear against a firmed transparent material, for instance glass. In [14] researchers suggest that the ear may have advantages over the face for biometric recognition. Their previous experiments with ear and face recognition, using the standard principal component analysis, showed lower recognition performance using ear images.

Although there are an appreciable number of methods that have been considered on ear biometrics [15], [16], no one yet has a sufficient performance. A main disadvantage that appears to all ear based applications makes their popularity even lower. The subject has to turn its head perpendicularly to the camera or even worse to stick its ear in a board in order the sample to be taken. A relevant security system would be considerably intrusive.

## 2.4 Nail ID

A really novel and eccentric idea is presented by AIMS Technology Inc. [17]. The company released a commercial product that is supposed to identify a person by reading the information that is hidden in finger nail and more specifically in nailbed. The nailbed is an essentially parallel epidermal structure located directly beneath the fingernail. The epidermal network beneath the nail is mimicked on the outer-surface of the nail. Rotating one's fingernail under a light reveals parallel lines spaced at intervals. The human nailbed is a unique longitudinal, tongue-in-groove spatial arrangement of papillary papillae and skin folds arranged in parallel rows.

Keratin microfibrils within the nailbed are located at the interface of the nailbed and the nailplate, or fingernail. AIMS relies on widely used technology and known principals. Essentially, AIMS utilizes a broadband interferometer technique to detect polarized phase changes in back-scattered light introduced through the nailplate and into the birefringent cell layer. By detecting the phase of the polarized optical signals corresponding maximum amplitude, one can reconstruct the nailbed dimensions using a pattern recognition algorithm in the interferometry. The identification process generates an one-dimensional map of the nailbed, a numerical string much like a "barcode" which is unique to each individual. AIMS believes that their design track will, at the end of the development cycle, result in an in-expensive hardware scanning assembly.

In the question regarding the effectiveness and efficiency of this technology, producers reply that this technique is likely to be more effective and efficient than other verifiers. The nailbed, residing beneath the nailplate, is not externally visible and hence difficult to alter or duplicate.

The Company provides many possible advantages that would result from the use of the nail scanner with most important the higher level of confidence. It is doubtless that all of the specifications given for the “Nail ID” sound very interesting. However, information given derives exclusively from internal sources. So far, there is no kind of published work showing the true capabilities of the system.

## 2.5 Otoacoustic Emissions Recognition

A new way to confirm a person’s identity when using a credit card is by using otoacoustic emissions. A research project at the University of Southampton is examining whether hearing could be effective in recognizing individuals by this method [18]. The researchers are examining the reliability of using this as a distinguishing feature.

One application might be to guard against mobile phone theft, where an acoustic system could check on whether the user matched the profile of the owner. Scientists at the University of Southampton report that a routine test carried out on infants to detect possible hearing problems may also be used to distinguish between individual, mature adults. They maintain too, that if clicks are broadcasted into the human ear, a healthy ear will send a response back [18], [19]. These are called otoacoustic emissions. For instance this could mean some kind of telephone receiver being used for card transactions, presumably in conjunction with a PIN number. A cardholder would pick up the receiver and listen to a series of clicks and the response would be measured and checked against the information stored on the card and the records held by the credit card company or bank.

It is worth noticing that from the total of 704 measurements in [18], 570 (81%) were identically classified. The results seem to be low for a confident and secure real word application. It is very promising though, for a technology that makes its first steps in a completely new aspect of biometrics.

## 2.6 Skin Spectroscopy

In [20] a new biometric technology based on the unique spectral properties of human skin is described. Skin is a complex organ made of multiple layers, various mixtures of chemicals and distinct structures such as hair follicles, sweat glands and capillary beds. While every person has skin, each person’s skin is unique. Skin layers vary in thickness, interfaces between skin layers have different undulations and other characteristics, collagen fibres and elastic fibres in the skin layer differ and capillary bed density and location differ. Cell size and density within the skin layers, as well as in the chemical makeup of these layers, also vary from person to person.

The system’s hardware and software as it is advertised recognize these skin differences and the optical effects they produce. The developed sensor illuminates a small (0.4 inch diameter) patch of skin with multiple wavelengths (“colours”) of visible and near infrared light. The light that is reflected back after being scattered in the skin is then measured for each of the wavelengths. The changes to the light as it passes through the skin are analyzed and processed to extract a characteristic optical pattern that is then compared to the pattern on record or stored in the device to provide a biometric authorization.

The technology developed from this particular company as it is presented, is avowedly very promising. The results

however, cannot be cross-checked with other researchers as details of the method used, are not provided. Company has its own reasons to keep this technology obscure and definitely to publish only what defends the robustness of the system. Reasonably then we have to keep a neutral attitude until the first experiments from an unprejudiced researcher will be published.

## 3. CONCLUSION

In this paper we introduced the latest achievements on biometric technology. All of the techniques proposed, according to the level of disturbance that they occur to the subject, are classified to intrusive or non-intrusive. Every method provides its own advantages compared to others already existed. However it seems to be too early to evaluate their capabilities properly.

## 4. REFERENCES


1. Diego A. Socolinsky, Andrea Selinger, and Joshua D. Neusheisel. Face recognition with visible and thermal infrared imagery. Computer Vision and Image understanding 91 (2003) 72-114.

2. Aglika Gyaourova, George Bebis and Ioannis Pavlidis. Fusion of Infrared and Visible Images for face recognition. European Conference on Computer Vision (ECCV), Prague, May 11-14, 2004.

3. E. Guan, Sara Rafailovich-Sokolov, Isabelle Afriat, Miriam Rafailovich and Richard Clark; Analysis of the Facial Motion Using Digital Image Speckle Correlation. Materials Science and Engineering, State University of New York; Stella K. Abraham; Biomedical Engineering, Mechanical Properties of Bio-Inspired and Biological Materials V. MRS fall Meeting.

4. Makoto Omata, Takayuki Hamamoto, and Seiichiro Hangai. J. Bigun and F. Smeraldi. Lip Recognition Using Morphological Pattern Spectrum (Eds.): AVBPA 2001, LNCS 2091, pp. 108-114, 2001. Springer-Verlag Berlin Heidelberg 2001.

5. Zhihong Pan, Glenn Healey, Manish Prasad, and Bruce Tromberg. Face Recognition in Hyperspectral Images IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 12, December 2003.

6. M. Ohlsson, H. Holst, L. Edenbrandt, Acute myocardial infarction: analysis of the ECG using artificial neural networks, in: Artificial Neural Networks in Medicine and Biology (ANNIMAB-1), Goteborg, Sweeden, 2000, pp. 209–214.

7. L. Biel, O. Pettersson, L. Philipson, P. Wide, ECG analysis: a new approach in human identification, IEEE Trans. Instrum. Meas. 50 (3) (2001) 808–812.

8. R. Hoekema, G.J.H. Uijen, A. van Oosterom. Geometrical aspect of the interindividual variability of multilead ECG recordings, IEEE Trans. Biomed. Eng. 48 (2001) 551–559.

9. StevenA. Israela, John M. Irvineb, Andrew Chengb, Mark D.Wiederholdc, Brenda K.Wiederhold. ECG to identify individuals. 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved. doi:10.1016/j.patcog.2004.05.014

10. Artificial intelligence to increase security of online shopping and banking Queen Mary, University of London 29 August 2003 ©. http://www.qmw.ac.uk/poffice/nr270803.shtml.

11. Naoto Miura, Akio Nagasaka, Takafumi Miyatake. Feature extraction of finger-vein patterns based on iterative line tracking and its application to personal identification. Systems and Computers in Japan, Vol. 35, No. 7, 2004.

12. SecuaVeinAttestor. Hitachi Engineering Co. Ltd. http://www.hitachi-hec.co.jp/virsecur/secua_vein/vein01.htm

13. M. Burge and W. Burger. Ear Biometrics for Machine Vision. Ear Biometrics in Computer Vision, in the 15th comference of the Pattern Recognition, ICPR 2000, p. 826-830.

14. Kyong Chang, Kevin W. Bowyer, Sudeep Sarkar, and Barnabas Victor. IEEE Comparison and Combination of Ear and Face. Images in Appearance-Based Biometrics, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, NO. 9, September 2003.

15. Hurley, D.J., Nixon, M.S., Carter, J.N. A New Force Field Transform for Ear and Face Recognition. In Proceedings of the IEEE 2000 International Conference on Image Processin ICIP 2000b, pp. 25-28.

16. Moreno, B., Sánchez, Á., Vélez. J.F. On the Use of Outer Ear Images for Personal Identification in Security Applications. IEEE 33rd Annual International Carnahan Conference on Security Technology, 1999, pp. 469-476.

17. Nail ID 843-399-2202. AIMS Technology Inc. Biometrics Systems Division http://www.nail-id.com/index.html.

18. Hoppe, U., Weiss, S., Stewart, R. W. and Eysholdt, U. (2001) An Automatic Sequential Recognition Method for Cortical Auditory Evoked Potentials. IEEE Transactions on Biomedical Engineering 48(2):pp. 154-164.

19. Dietl, H. and Weiss, S. Cochlear Hearing Loss Detection System Based on Transient Evoked Otoacoustic Emissions. (2004) In Proceedings of IEEE EMBSS Postgraduate Conference, Southampton.

20. Robert K. Rowe, Stephen P. Corcoran, Kristin Nixon. Biometric Identity Determination using Skin Spectroscopy Lumidigm, Inc., 800 Bradbury SE, Suite 213, Albuquerque, NM, USA 87106 www.lumidigm.com.

## SYSTEMS AND APPLICATIONS

# SAGENT: A MODEL FOR EXPLOITING BIOMETRIC BASED SECURITY FOR DISTRIBUTED MULTIMEDIA DOCUMENTS

*G. Howells, H.Selim, M.C.Fairhurst, F.Deravi and S.Hoque*

Department of Electronics, University of Kent, Canterbury, Kent UK
E-mail: W.G.J.Howells, M.C.Fairhurst, F.Deravi, S.Hoque@kent.ac.uk

## ABSTRACT

Issues concerning the security of heterogeneous documents in a distributed environment are addressed by introducing a novel document model capable of incorporating any desired, and in particular biometrically based, security techniques within existing documents without compromising the original document structure.

.

## 1. INTRODUCTION

In the modern distributed data processing environment, security issues regarding the integrity of documents are of particular concern. This paper describes the Securable Autonomous GENeralised documenT model (SAGENT), a document authoring and management system capable of addressing the significant problem of incorporating any given security algorithm within a distributed, heterogeneous document. The purpose of the current paper is to focus on illustrating how a document security mechanism based on biometric encryption may be applied to documents developed within such a model rather than document structure which was introduced in [6].

The model allows:-

- The employment of object-oriented concepts, (encapsulation, multiple inheritance etc. [1])
- Language independence, allowing a document to incorporate methods written using an arbitrary programming language.
- Flexibility to accommodate different document types including potential future innovations.
- Additions to the model to be incorporated without compromising its original structure.
- Provision for whole or selective encryption (and data compression), possibly multi-layered encryption schemes and support for authentication measures and biometric identification schemes [2].

## 2. SAGENT MODEL OVERVIEW

A primary motivation for the SAGENT model is to deliver improved security suitable for employment in a modern distributed environment. To achieve this aim, the model seeks to address a number of key issues affecting the performance of existing systems. A significant issue is that of *autonomy*. The SAGENT model allows for all functions for the manipulation and access control of the document to themselves form an integral part of the virtual distributed document whilst also allowing these components to reside within distributed locations. The complete document can therefore be treated as an autonomous entity because it not only holds the data but also the basic structure and the associated operations required to manipulate the document. This makes the model extremely robust and flexible although it does allow for key components such as compilers and encryption algorithms to be shared between documents to avoid unnecessary physical duplication.

A further key issue is that of *generality.* The SAGENT model seeks to provide a system which is capable of manipulating and enhancing the capabilities of existing document models, allowing their features to be exploited unhindered, whilst simultaneously allowing additional further desirable features such as those associated with security and encryption. Generality requires that these algorithms be expressible in any arbitrary format required by the document authors.

To achieve these aims, it is necessary to provide a mechanism capable of seamlessly incorporating the features of existing models within an enhanced model definition. Ideas associated with object-oriented modelling [3] have become well established and are able to provide a suitable basis for the SAGENT model definition. Each SAGENT document is abstractly represented as a series of objects with each object being a member of a given class. The classes themselves, when taken together, form a hierarchy. Importantly, all information and algorithms for accessing and processing the document may be contained within properties and methods associated with the document objects themselves. Although this does not provide an absolute guarantee of integrity when executed on arbitrary platforms and within arbitrary network environments, nevertheless, by involving such environments to a minimum in its operation, SAGENT minimises the ability of such platforms and environments to compromise its integrity. The root of the system is a class, named *Root,* over which are defined properties and methods relating to all document objects. The additional concept of *multiple inheritance* allows multimedia documents to be modelled as a single entity.

The SAGENT model has been designed to allow the component objects of the document to reside at different physical locations. This is important to avoid the problem of large physical documents which require many support methods in order to efficiently operate. Essentially, a SAGENT document is one which is aware of the locations of secure and trusted methods which it may employ to perform its necessary access control and editing operations. Such

methods will be encrypted with the encryption model currently in use for the SAGENT document and thus not generally available outside the SAGENT system although such systems may be shared by various trusted SAGENT documents to avoid unnecessary physical duplication of data.

Additionally, the concept of *meta class* allows for dynamic modification of the class hierarchy and the methods within it. Essentially, each class within the hierarchy is itself represented as an object and the class of this object is defined to be the *Meta class* of the original class. The *Meta class* contains properties and methods describing the class (for example, defining its superclass and constituent methods). In the light of the above requirements, a language independent data-structure is proposed which implements the features of an object-oriented document. In theory, arbitrary authentication protocols may thus be incorporated within the SAGENT framework.

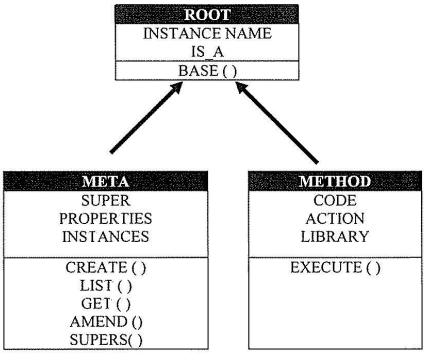



**Figure 1: Attributes and Methods for the Minimal Model**

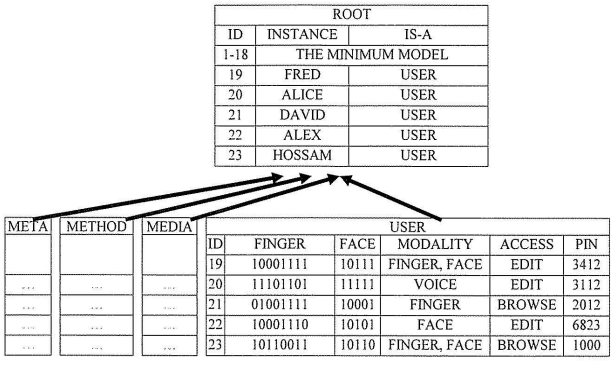
Figure 1 illustrates the *meta class* concept using the classes *Root*, *Method*, and their defining class *Meta*. The class hierarchy between *Root*, *Method* and *Meta* is maintained using one or more *class attributes* with the actual document objects and associated supporting information retained in the instances of these classes. Since *Meta* itself is a class like classes *Root* and *Method*, it is also represented as an object instance of *Meta*. The *Meta* class definition therefore becomes part of the class hierarchy itself [4].

Further to representing each class as an object, each method is also represented as an object, this object being an instance of a class *Method*. Methods defined over this class themselves return information about the various instances of the *Method* class (such as method name or method argument type) and are themselves also represented as instances of the *Method* class. The class hierarchy can be theoretically modified at any stage by the end-user, although in practice some restrictions may be applied according to the access requirement of a given document [5]. It should be emphasised that this proposal does not compromise the ideal of modular software development nor the modern practice of separating data and presentation information. It merely adds an additional envelope layer for all logical components of the document in order to implement arbitrary access and security protocols. There is, however, obviously an overhead associated with the production and maintenance of the meta data associated with the SAGENT model. It should be noted however that such meta data must be represented by some mechanism in any alternative model. Within the SAGENT model, the overhead is consistent with that which may be found within any object oriented database system and although alternative mechanisms may provide for increased efficiency in data retrieval, the flexibility offered by the object oriented paradigm often outweighs any such benefits. The flexibility of meta data representation is a significant advantage of the SAGENT system. The model has been designed to require only the minimal amount of meta data as an initial starting point for model construction. The classes *Root, Meta and Method*, together with their associated properties and methods, form the *Minimal Model* of the SAGENT system which must be present for any SAGENT document to be accessed correctly. Apart from the Minimal Model, SAGENT documents are free to expand to incorporate any features, algorithms or access protocols that an author requires. A major goal of the SAGENT model is to ensure that the Minimal Model strictly contains only the minimum that is required in order to allow maximum flexibility in document design. The constituents (*attributes* and *methods*) of the individual minimal model classes are shown in Figure 1.

It should be emphasised that although, in order to allow maximum generality, the minimal model is the theoretic starting point for any new document, in practice, it is envisaged that templates will be available for standard document forms such as XML and MPEG-7 based documents. These templates will contain additional classes and instances to address the requirements of the standard being modelled. Further details of the SAGENT model may be found in [6].

# 3. BIOMETRIC SECURITY

A major goal of the SAGENT model, and the principal focus of this paper, is the efficient integration of biometric measurements as a means of determining and monitoring personal identity in order to control and manage access to virtual documents in circulation within a user community. The following is a description of how biometric identification may be incorporated within a SAGENT document. The *generality* of the SAGENT system, however, means that many other schemes are possible to meet particular user requirements.



ROOT

| ID | INSTANCE | IS-A |
|---|---|---|
| 1-18 | THE MINIMUM MODEL | |
| 19 | FRED | USER |
| 20 | ALICE | USER |
| 21 | DAVID | USER |
| 22 | ALEX | USER |
| 23 | HOSSAM | USER |

| META | METHOD | MEDIA |
|---|---|---|
| | | |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |

USER

| ID | FINGER | FACE | MODALITY | ACCESS | PIN |
|---|---|---|---|---|---|
| 19 | 10001111 | 10111 | FINGER, FACE | EDIT | 3412 |
| 20 | 11101101 | 11111 | VOICE | EDIT | 3112 |
| 21 | 01001111 | 10001 | FINGER | BROWSE | 2012 |
| 22 | 10001110 | 10101 | FACE | EDIT | 6823 |
| 23 | 10110011 | 10110 | FINGER, FACE | BROWSE | 1000 |

**Figure 2:** The incorporation of biometric data in the model

Figure 2 shows how biometric measurements may be incorporated as a means to verify the identity of users by creating a new class named *User*. *User* class represents the

users within the model, each user possessing a *name* and other biometric information that verifies the identity of the user. In the example presented, two biometric measures are adopted representing data extracted from fingerprint and facial image features. The additional class *Media* is assumed to contain the actual data associated with the document content whose precise structure is not of interest in the current context.
The details of the *User* class are as follows:-

*Attributes*

- *Fingerprint*: holds the string of bits representing the fingerprint images or extracted features.
- *Face*: holds the data for facial image features.
- *Pin:* holds the pin or password for each user
- *Modality:* holds the *name* of the biometrics required for the user to gain access to a document.
- *Access Category*: holds an indicator of the level of permitted access allocated to the user (*edit, browse, append*)

It should be noted that *Fingerprint* and *Face* together form two example attributes of biometric data and their inclusion does not imply that these examples are superior to other form of biometric identification or that they are the only form of biometric technique applicable to the SAGENT system. It should also be noted that the storage of such template information for biometric data, and any unauthorised access to them, would represent a serious breach of security as other documents and systems protected by the biometrics would also be compromised. There are two points to note regarding this:-

- The SAGENT system may encrypt all template data using arbitrary encryption algorithms. SAGENT is thus as secure as any other template based biometric authentication system.
- Alternative template-free biometric systems involving the direct generation of encryption keys from the biometric data are fully compatible with the SAGENT model. Although such systems are ,at present, at an early stage of development [7], the generality property of the SAGENT system will allow them to be cleanly incorporated within SAGENT documents when they become available.

*Methods*

- *addUser*: to allow the creation of a new user: this method is available to the document owner only.
- *setUserCategory*: allows for the modification of the access category assigned to a user to access the document resources. This method is again available to document owners only.
- *Finger enrolment:* fetches the sample fingerprint data for the claimed user.
- *Finger verification:* verifies the sample fingerprint data .
- *Face enrolment:* fetches the sample face data for the given user
- *Face verification*: verifies the sample face data
- *Pin Verification*: verifies the Personal Identification Number provided by user

When a new document is created, the creator of the document is referred to as the document *Owner* and is able to modify the *User* class. Unless authorised to do so by the *Owner*, no user subsequently created will be authorised to change the user class itself. Only the *Owner* will normally possess such a privilege. The model enforces this condition by checking the biometrics of this *Owner*. The enforcement checking is done by using a *Verification* method in the *User* class which is invoked when the subject sends a request to access the document.

Initially, no further users are present and therefore there is only a basic document whose unique user is the *Owner*. Once the document has been created, the Owner is able to include the users of the application using the method *addUser*. Subsequently, the *amend* method may be employed to assign data to verify the identity of a user. The *Amend* method is part of the Minimal Model and is valid for all the attributes in the *User* class except the *category* attribute. The *category* attribute can be amended by the *setUserCategory* method The final step involves the assignment of the appropriate access category to each user by means of the *setUserCategory* method. Users may then enhance the structure of the document and include their own contents and other required elements which may already be contained within other SAGENT documents.

It should again be emphasised that the above scenario is an example of an access control system for incorporating biometric identification. Typically, once verified, access will be given to an encryption private key which the user may then use to access the data comprising the model. Alternative techniques may be incorporated via the inclusion of appropriate classes and methods within the model.

# 4. BIOMETRIC VERIFICATION

This section introduces the detail of the biometric verification process described above. The system is designed to allow the maximum flexibility in the access requirements for each user whilst retaining the security integrity of the system.

## 4.1 User Verification

Consider again Figure 2. Here, the user may access the secure document object after the verification performed by the biometric authentication method. Each user provides a *name* followed by the live biometric data in accordance with the value of the *Modality* attribute for each user. The verification process proceeds as follows:-

- The model retrieves the instance representing the user from the *user* class using its *user identifier or name.* The details of the *retrieve* process are described in [6].
- The value of the *Modality* attribute is retrieved from this instance which lists the biometric methods required for the user to gain access to the document.
- Each value in the *Modality* attribute represents one or more instances in the *Method* class. The result of each biometric method given in the *Modality* attribute is returned to the verification method. In the case of acceptance, this will generate a decryption key which will allow the user access to the permitted areas of the document as described below.

## 4.2 Decryption of the Document Data

Although the SAGENT model does not require the data contained within a document to be encrypted, typical secure

usage of the system will normally employ such techniques. The generality principle underlying the SAGENT model allows for the flexibility of employing any encryption algorithm desired by an author, since the relevant function needs merely to be incorporated within the meta data. This section presents on overview of two alternatives for associating the encryption keys with the biometric verification system presented above.

The most obvious strategy for synergising biometric verification and encryption is to release the decryption key when the biometric identification of the user has been verified. This would mean the decryption key is released by the verification methods on successful authentication. The drawback to this is that the key itself is stored within the meta data of the model and may be discovered (although this is not a trivial task) by informed analysis of the raw document data.
A far stronger alternative is to generate asymmetric encryption key pairs directly from the biometric data supplied by the user. In this scenario, rather than hold template biometric data, the verification methods (for example *Finger verification* and *face verification)* will generate the decryption keys directly from the user biometric data supplied. The difficulties here are to ensure that a biometric sample produces stable keys whilst simultaneously ensuring a given user possesses a unique key. Significant work has been undertaken by the authors and others in investigating techniques for achieving this aim [7]. As stated above, such techniques alleviate the risks associated with explicitly storing biometric template data. The major point relating to this paper, however, is that the SAGENT model provides a general, flexible and powerful framework for incorporating such techniques as and when they become widely available.

## 4.3 Verification Example

Again referring to Figure 2, consider the scenario where an arbitrary user *Fred* wishes to access a document. The process introduced above is employed. Initially the user will enter the name by sending the following message to the document:

*Verifiy ( Fred )*

The model will perform the following steps: -

- The object which has the identifier corresponding to the user *Fred*, in this example *19*, will be retrieved from the *User* class. Referring to Figure 2, note that it possesses the following attribute values:-

| | | | | | |
|---|---|---|---|---|---|
| 19 | 10001111 | 10111 | Finger, face | edit | 3412 |

- The *verification* method will retrieve the *modality* attribute by sending the message

  *19 . Get (Modality)*

  The result "*finger* and *face*" will be returned implying that the user will access the document after the verification of the two biometric methods *finger* and *face*.
- The *verification* method will subsequently send the following message

  *Execute (finger verification, <finger data>)*

  The *finger verification* is an instance in the *Method* class and the result of this method will be returned to the verification method. If the result is accepted, the following value in the modality attribute will be executed by sending the message

  *Execute (face, <face data>)*

  The final decision to allow the user to access the document is taken according to the result of the two methods *finger* and *face* and access to the document will be granted via the encryption techniques outlined in section 4.2.

# 5. CONCLUSION

Significant problems associated with the security of documents in a distributed environment have been addressed. A powerful and flexible model has been introduced allowing:-

- The incorporation of arbitrary security features allowing the latest security techniques to be efficiently included within existing standard document definitions
- Document autonomy by incorporating all relevant document processing features within the distributed virtual structure of the document. This removes the dependence on external systems and ensures all such features are encompassed by the access and encryption features of the document whilst not requiring such features to reside at the same physical location or to be unnecessarily duplicated.
- Efficient incorporation of biometric security features allowing for direct utilisation of biometric technology within the encryption systems of the document.

Overall, these features offer the potential of a significant increase and timely improvement in the security and integrity of distributed documents at a time when such issues are of particular concern and without compromising modern modular document development standards or techniques.

# 6. REFERENCES


1. M. Bouzeghoub, G. Gardarin and P. Valduriez, "Object technology, concepts and methods", Boston, London, International Thomson Computer Press, 1997
2. W. Shen, and R. Khanna, "Automated Biometrics", *Proceedings of IEEE*, pp. 1343-46, September 1997.
3. R. Binder, " Testing object-oriented systems, models, patterns, and tools, " Reading, Mass., Harlow, Addison-Wesley, 2000
4. A. Sheth, and W. Klas, "Multimedia data management, using metadata to integrate and apply digital media" New York, London, McGraw-Hill, 1998.
5. F. Deravi, "Audio-Visual Person Recognition for Security and Access Control", JTAP Report 38, JISC, September 1999.
6. G. Howells, H. Selim, S. Hoque, M.C Fairhurst M.C, and F. Deravi, "The Autonomous Document Object (ADO) Model," *proceedings of 6th International Conference on Document Analysis and Recognition (ICDAR),* Seattle, Washington, 2001.
7. Y. Yamazaki and N. Komatsu. A Secure Communication System Using Biometric Identity Verification. *IEICE Transactionon Inf. And Syst.* **E84-D**(7):879-884. July 2001.

# WAVELET-BASED FACE VERIFICATION FOR MOBILE PERSONAL DEVICES

*Sabah Jassim, Harin Sellahewa, & Johan-Hendrik Ehlers*

Department of Information Systems
University of Buckingham, Buckingham MK18 1EG, U.K.
Email: {sabah.jassim;harin.sellahewa;johan-hendrik.ehlers}@buckingham.ac.uk

## ABSTRACT

The fast growing ownership and use of mobile personal programmable communication devices (e.g. PDA's) create new and exciting opportunities for convenient and secure commercial and service providing transactions. The latest generations of PDA's incorporate digital cameras and signature pads, raising the possibility of enhancing the security of mobile transactions using biometric-based authentication, but they are constrained in their memory size and computational power. This paper is concerned with wavelet techniques for face verification as part of a multi-modal biometrics-based identity verifier for PDA's that would include voice and hand-written signature. We report on the performance of our scheme, in comparison to known face schemes, using the BANCA and the ORL databases.

**Keywords:** Biometrics, SecurePhone, PDA and Wavelets.

## 1. INTRODUCTION

Human Identification based on facial images is one of the most challenging tasks in comparison to identification based on other rather obtrusive biometric features such as fingerprints, palm-prints or iris [1]. For biometrics verification on PDA's to be acceptable to users, efficiency becomes an essential requirement. This is a very tough challenge due to the constrained memory and computation power of such devices. On such constrained devices, it is not realistic to apply some/all of the pre-processing procedures that are used for most automatic face verification such as normalising the effect of variations in recording conditions. SecurePhone is a European Union (EU) funded project that aims to develop a multi-modal biometric verifier for PDA's, using voice, face and handwritten signature. In this paper, we are concerned with face verification (i.e. the one-to-one case) component. We develop a wavelet-based face verification scheme suitable for implementation on currently available PDA's with acceptable accuracy rate. We shall also test the performance of our scheme using the BANCA and the ORL databases.

Face verification involves 4 distinct stages, each with its inherent difficulties: Face location, Face normalisation, Feature extraction, & classification. *Face location* is a difficult task if there are many different background objects [1], [2], [3]. *Normalisation* transforms the located face image into a normal form in terms of size, pose, expressions, rotation and illumination [1], [4]. Performance of a face verification/ recognition system is affected by variations in above conditions [1], [2]. – A typical face image is represented by a high dimensional array (e.g. 120x120). It is difficult to compare two face images as they are (original images) since the images are too large and they contain areas, which are similar in most faces [2] (i.e. cheeks, forehead). This could have an adverse effect on the verification accuracy. *Feature extraction* aims to represent a face image in a much lower dimensional **feature space** that encapsulates the essential facial information. Two current approaches to feature extraction are the geometry feature-based methods and template-based methods [1], [2], [4], [5], [6]. In the latter more common approach the entire face image is statistically analysed to obtain a set of feature vectors that best describe a given face image. Face images are first linearly transformed into a low dimensional subspace and then represented as compact feature vector in this subspace. Typical dimension reduction methods are based on Principle Component Analysis (PCA) [1], [2], Linear Discriminate Analysis (LDA) [5], [6], [7], Independent Component Analysis (ICA) [1] or a combination of these methods. *Classification* is the decision stage of a verification/recognition scheme. When a new face image is given, its feature vector is obtained and compared with the enrolled image set using a similarity/distance function. Typical methods used are City Block (L1), Euclidean (L2), Correlation, Covariance, Mahalanobis distances or SVM (Support Vector Machines) [4].

## 2. PROGRAMMABLE MOBILE PHONES

Currently available mobile communication devices (3G smart phones) and Personal Digital Assistant (PDA's) are equipped with a camera which can capture both still and streaming video clips and a touch sensitive display panel. A SIM card is a microprocessor, which can securely store sensitive data (e.g. subscriber details, private keys and other user data). PDA's are available with processing speeds of up to 400MHz and 128MB of RAM [8]. Current Subscriber Identity Module cards (SIM) have a memory capacity of up to 256K. Beside convenience, PDA's can provide better secure infrastructure for financial or sensitive transactions than ATM and online transactions, and protect against fraud and repudiation while allowing for assigning clear responsibilities for such threats.

Most existing face verification schemes require relatively large amounts of data to be stored for the feature extraction process [9]. The necessary computational resources are far beyond the modest capabilities of available PDA's. For the proposed SecurePhone project application, it is vital that all enrolment data of the 3 biometric modalities are within the available storage capacity on a SIM card, although it is possible to use the not-so secure PDA storage. Here we propose an efficient

feature representation method for facial images based on wavelet transforms. On the other hand, limited processing power and speed of communication between the PDA and the SIM means that we have to reduce the amount of preprocessing to the minimum. On PDA's most resources should be dedicated to feature extraction and classification rather than to face location or normalisation. The size and pose normalisation can be avoided by placing the face inside a fixed, reasonably sized box, whereas variation in lighting conditions must be dealt with in accordance with the normal indoor/outdoor varied use of PDA.

## 3. WAVELET TRANSFORMS

The wavelet transform is a technique for analyzing finite-energy signals at multi-resolutions. It provides an alternative tool for short time analysis of quasi-stationary signals, such as speech and image signals, in contrast to the traditional short-time Fourier transform. The Discrete Wavelet Transform (DWT) is a special case of the WT that provides a compact representation of a signal in time and frequency that can be computed efficiently. The DWT decomposes a signal into frequency subbands at different scales, from which the signal can be perfectly reconstructed. The mathematical properties of the DWT is equivalent to filtering the input image with a bank of band-pass filters whose impulse responses are approximated by different scales of the same ***mother wavelet***. It allows the decomposition of a signal by successive highpass and lowpass filtering of the time domain signal respectively, after sub-sampling by two.

There are a number of different ways of applying a 2D-wavelet transform. The most commonly used wavelet decomposition of an image, that we adopt here, is the *pyramid* scheme. At a resolution depth of k, this scheme decomposes an image I into 3k +1 subbands, $\{LL_k, HL_k, LH_k, HH_k,\ldots,HL_1, LH_1, HH_1\}$, with $LL_k$, being the lowest-pass subband. The subbands $LH_1$ and $HL_1$, contain finest scale wavelet coefficients that get coarser with $LL_k$ being the coarsest, (see Figure 1). The $LL_k$ subband is considered as the k-level approximation of I.

Recently, wavelet transforms have been used for face recognition, mostly combined with LDA schemes [6], [10], [11]. Reducing image dimensions using wavelet decomposition and then applying PCA or other feature extraction methods on the low subband image coefficients can further reduce computations (smaller covariance matrixes, small and less number of eigenfaces) and storage requirements. Recent dimension reduction approaches to face recognition include downsizing images by averaging, rankles and waveletfaces [6], [12]. Previously, we applied PCA to the $LL_k$ of facial images down to level k = 5 and the performance of these systems were comparable, if not better than, PCA in the spatial domain [13]. We have also shown that using the LL-subband itself as the face feature vector results in comparable or even higher accuracy rate than PCA in the spatial or the wavelet domain. The computational efficiency of evaluating and comparing the LL-subband at the enrolment and classification stages, make it suitable for constrained devices. The non-LL subbands do hold information that can be used to improve the performance of the LL-scheme and are needed for face location and feature-preserving compression required for online face verification [14]. In the rest of the paper, we focus on on-the-device face verification.

## 4. WAVELET DOMAIN FEATURE VECTORS

Our feature representation of a face image I is based on the LL-subband of the wavelet decomposed image WT(I). Different decomposition level and/or wavelet filters yield different representations. The depth of decomposition is chosen according to efficiency requirement and the size of the face in I. There are many different wavelet filter banks to use in the transformation stage, and it is possible that the choice of the filter may influence the accuracy of the recognition / verification scheme. However, in this paper we only report on the use of the Haar and the Antonini filters. The use of the Daubachie 4 filter did not show a significant difference to the Haar filter. Our previous experimental results show a significant improvement in performance when we standardized the LL coefficients.

The significance of wavelet-based schemes for constrained devices stems from the fact that the size of the k-th low subband is $1/4^k$ of the original image size. It is worth noting that each wavelet coefficient in the $LL_k$ subband is a function of a $2^k x 2^k$ block of pixels in the original image representing the scaling of the total energy in the block scaled by the given mother wavelet. Moreover, working at wavelets beyond subband 3 is expected to result in robustness against high rate compression or noise interference.

## 5. EXPERIMENTAL RESULTS & DISCUSSION

To evaluate the performance of our verification scheme, we conducted experiments on the controlled section of BANCA and the ORL database. In [13], we reported on experiments that tested the performance of various wavelet-based face verification schemes on the non-video degraded version of BANCA, to reflect the rather low quality of recordings done on current PDA's. Both databases include a lot of variations in size, pose, head orientation and illumination. For PDA's we can expect a certain level of cooperation from the user (e.g. willingness to place his/her face in a predefined area of the display panel in a near frontal pose). To test the effect of variation in illumination, we test the performance of our scheme with and without Histogram Equalisation.

The BANCA database, [15], is a benchmark for testing the performance of biometric schemes. It consists of a number of sections. The English part of the database consists of 52 subjects with equal number of females and males. For each subject, there are 4 recording sessions, indexed from 1 to 4. Each session consists of a true client video clip where the subject utters own information, and another video clip while uttering details of another subject. The face areas were cropped semi-manually and then resized to a uniformed size of 160x192.

In the non-video BANCA tests, the best performances were achieved when the LL4-subband of Haar transformed images or the LL4-subband Antonini were used as feature vectors. Therefore, in the new set of experiments on the video version of BANCA, we only used these two filters. The face areas in the video frames

were wavelet transformed to depth 4 of decomposition, using the Haar as well as the Antonini wavelet filters. For each cost ratio R, we tested the performance of 4 variants of the wavelet scheme: LL4-subband, standardised LL4-subband, LL4-subband of the Histogram Equalisation (HE) images, and the standardised LL4-subband of the HE images. The HE procedure has been used by many existing face verification schemes as a mean of reducing the effect of variation in lighting conditions. The results from these experiments, shown in table 1, demonstrate highly improved WER performance (for three cost ratio values) as compared to the previously reported performance of the baseline scheme on the non-video version of the degraded section of BANCA, and they are comparable with the top performing state-of-the-art schemes. Table 2, include WER performance of the best face verification scheme obtained in the ICBA – competition (CSU software), [16]. Table 2, also shows that the application of HE before the WT improved the performance across all cost ratios.

The ORL database consists of face images for 40 subjects, each with 10 facial images of 92 x 112. We cropped the face areas to 80 x 96 pixels, removing only background areas whenever possible. The images were wavelet transformed each to level 3, to get 10x12 LL3-coefficients as the feature vector. We then tested performance for two different feature vectors he LL3-only, LL3 after Histogram equalisation. For the WER calculations we divided the 40 subjects into 2 groups of 20 and the average WER for in the two groups are shown in table 2, below. As seen from these results, unlike the result for BANCA, HE led to degraded performance. Then we applied an automatic Gamma-transform, whose gamma value depends on the means of pixel values in a 2x2 block partition of the image. We are still developing this method. The improved performance is greater at R=0.1 and R=1.

Except for the wavelet-with-HE, the EER values outperform published results for ORL. The best performing scheme, reported in [17], has a 5% EER.

The results from these two sets of experiments show a significant discrepancy in the performance of our schemes (and indeed other schemes) over two databases. This may require close analysis of the nature of the two databases but we can observe that there are differences in the type of pose variation between the images in the two databases. Face orientation in ORL varies in the left-to-right direction, but in BANCA the variation is mostly in the up-to-down direction. In many BANCA videos, the subject is looking down obscuring their eyes.

In order to provide a more credible assessment of the performance and suitability of our scheme for use on smartphones, we need to use a PDA recorded database. Indeed, within the SecurePhone project, we are creating such a dedicated Audio-Visual database to be released shortly.

Finally, we conducted some tests to measure time taken by some of the image processing/transforming on the intended PDA (XDA II, also known as QTEK2020). For a cropped image of size 192x160, Haar to level 4 takes 0.33 sec, Antonini to level 4 takes 1.176 sec, and both HE and Gamma take 0.088 sec. The wavelet transforms can be optimised greatly, in particular if only LL-subband is required.

## 6. CONCLUSION

We have presented a wavelet based face verification/identification scheme as an efficient baseline scheme suitable for implementation on constrained devices such as PDA's and Smartcards. We compared its performance, under the further constraint of limited or no further prepossessing, with that a set of existing state of the art schemes. The results demonstrate a highly satisfactory performance in contrast to other computationally costly face verification schemes. Our tests demonstrate that in most cases, even without normalising recording conditions, the baseline scheme performs as well as the best well known PCA and PCA+LDA schemes for R=0.1 and R=10. For R=1, the performance is not as good but it is comparable with the average of those known schemes [4], [16]. Beside accuracy and efficiency, wavelet-only scheme require no training and the system can evolve with time. Whenever needed, new images can be added to the template.

Future work includes investigating the inclusion of a "world" model, adding statistical parameters from the non-LL subbands to the feature vectors, the use of automatic Gamma-transform before the wavelet. We are also investigating with other partners the use of the Gaussian Mixture Model (GMM), for fusing our scheme with the other proposed modalities within the SecurePhone project. Once, the PDA database is released we shall conduct further tests.

## 7. REFERENCES


[1] W. Zhao, R. Chellappa, A. Rosefeld, and P.J. Phillips. "Face Recognition: A Literature Survey," *Technical report, Computer Vision Lab, University of Maryland,* 2000.

[2] M. Turk, and A. Pentland. "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[3] D. Xi, and Seong-Whan Lee. "Face Detection and Facial Component Extraction by Wavelet Decomposition and Support Vector Machines," *Proc. AVBPA Int'l Conf. Audio-and Video-Based Biometric Person Authentication,* pp. 199-207, June, 2003.

[4] R. Beveridge, D Bolme, M. Teixeira, and B. Draper. "The CSU Face Identification Evaluation System User's Guide: Ver. 5.0," *Comp. Sci. Dept, Colorado State Uni.* (2003). http://www.cs.colostate.edu/evalfacerec/algorithms5.html, (10/12/2004)

[5] 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, Sept., 2000.

[6] Jen-Tzung Chien, and Chia-Chen Wu. "Discriminant Waveletfaces and Nearest Feature Classifiers for Face Recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 24, no. 12, pp. 1644-1649, December, 2002.

[7] Ming-Hsuan Yang. "Kernel Eigenfaces vs. Kernel Fisherfaces: Face Recognition Using Kernel Methods," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition,* pp. 215-220, May, 2002.

[8] D. Harfield, "Smartphone buyer," *Free with PDA Essentials*, issue 27, 2004, Highbury Entertainment Ltd.

[9] M. Sadeghi, J. Kittler, A. Kostin, and K. Messer, "A Comparative Study of Automatic Face Verification Algorithms on the BANCA Database," *Proc. AVBPA Int'l Conf. Audio-and Video-Based Biometric Person Authentication,* pp. 35-43, June, 2003.

[10] A. Z. Kouzani, F. He, and K. Sammut. "Wavelet Packet Face Representation and Recognition," *Proc IEEE Conf. Systems, Man, and Cybernetics*, pp. 1614-1619, 1997.

[11] Dao-Qing Dai, and P. C. Yuen. "Wavelet-Based 2-Parameter Regularized Discriminant Analysis for Face Recognition," *Proc. AVBPA,* pp. 137-144, June, 2003.

[12] F. Smeraldi. "A Nonparametric Approach to Face Detection Using Ranklets," *Proc. AVBPA,* pp. 351-359, June, 2003.

[13] ] H. Sellahewa and S. Jassim, "WAVELET-BASED Face Verification for constrained platforms", Proc. SPIE on Biometric Technology for Human Identification II, Florida 2005, Vol. 5779, pp 173-183.

[14] N. Al-Jawad and S. Jassim, "Feature-Preserving Image/Video Compression" To appear.

[15] E. Bailly-Bailliére, S. Bagnio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas. K Messer, V. Popovici, F. Porée, B. Ruiz, and J. Thiran. "The BANCA Database Evaluation Protocol," *Proc. AVBPA Int'l Conf. Audio-and Video-Based Biometric Person Authentication,* pp. 625-638, June, 2003.

[16] ICBA competition – CSU results. http://www.ee.surrey.ac.uk/banca/icba2004/csuresults.html (01/12/2004)

[17] Bicego et al., Probabilistic Face authentication using Hidden Markov Models, Proc. SPIE on Biometric Technology for Human Identification II, Florida 2005, Vol. 5779, pp 299-306.

## Tables And Images

| Verification Scheme | R=0.1 | R=1 | R=10 |
|---|---|---|---|
| **WT - Antonini** | 7.35 | 19.91 | 6.59 |
| **WT - Haar** | 8.35 | 23.15 | 5.83 |
| **std (WT) (Antonini)** | 8.96 | **16.24** | 4.82 |
| **std (WT)-Haar** | 10.24 | 18.24 | 6.07 |
| **WT (HE) -Antonini** | 7.35 | 17.21 | **4.59** |
| **WT (HE) - Haar** | 8.55 | 18.68 | 4.95 |
| **std (WT(HE)) -Antonini** | **7.06** | 16.75 | **4.59** |
| **std (WT(HE)) - Haar** | 8.63 | 17.69 | 4.83 |
| **Best of ICBA –CSU results** | **6.15** | **12.76** | **4.5** |

Table. 1

| WER For ORL | | | |
|---|---|---|---|
| **R** | **Image** | **Im+HE** | **Im+LGC** |
| **0.1** | 2.52 | 3.90 | 1.94 |
| **1** | 5.30 | 8.42 | 3.26 |
| **10** | 1.21 | 2.11 | 1.18 |
| **EER** | 3.62 | 6.5 | 3.38 |

Table. 2

**Figure 1**. From left – original dark, histogram equalised, auto gamma, Haar level 1, Haar level 3, and Haar after Gamma.

# A DISTRIBUTED MULTIMODAL BIOMETRIC AUTHENTICATION FRAMEWORK

*Jonas Richiardi, Andrzej Drygajlo, Alicia Palacios-Venin, Razvan Ludvig, Olivier Genton, Lianick Houmgny*

Perceptual Artificial Intelligence Laboratory
Signal Processing Institute
Swiss Federal Institute of Technology Lausanne
<jonas DOT richiardi AT epfl.ch>

**ABSTRACT**

The design and implementation of a distributed multimodal biometric verification system inspired by the BioAPI specification is exposed. Server and Client objects divide processing responsibilities in classes linked by the Java Remote Method Invocation framework. Our framework is secured up to the core security requirements of the X9.84 standard. It is designed to be extensible to other modalities than speech and on-line signature through object-oriented design.

## 1. INTRODUCTION

As biometric matchers have matured significantly in the past few years and their performance has been improved substantially for many modalities, there is now a need to explore practical implementations and establish the bases of software engineering practices and designs which will serve for integration of biometric software modules. This somewhat imprecise term can take several meanings.

One is that the biometric module is a stand-alone executable, which is just given a list of data files for training and testing, and produces matching scores as an output. This is often the case in competitions; for example during the Signature Verification Competition 2004 [1] the participants were asked to submit binaries conforming to a certain command-line syntax for training and testing.

Another interpretation refers to executables that perform part of the whole biometric processing chain (feature extractors, modelling engines, matching engines,...). An example is the `HList` program, part of the `HTK` toolkit [2], which can be used to extract features from speech files.

Yet another interpretation refers to generic pattern recognition/machine learning source code or toolkits in a particular programming language that can be used and applied to the biometric case. Examples abound, from C++ based toolkits such as ALIZE [3] to the Matlab Netlab library [4].

Standardisation efforts have matured largely in the industrial context, with the appearance of industry standards such as BioAPI [5, 6] and X9.84 [7], and more effort under way at the ISO level (ISO/IEC JRC 1/SC 37 series). Because we feel that most biometric researchers are rightly more interested in focussing their efforts on improving classifiers and combination methods than in following industry standards, we propose a flexible framework (the Multimodal Biometric Authentication Framework or MBAF) which can be used to experiment with biometric authentication in a distributed, secure system. It offers a large amount of flexibility in integration of existing biometric software modules and provides remoting, persistence, security, GUI, and workflow services. We believe this to be the first freely available distributed object framework of the kind, and it was used at the BioSecure residential workshop in Paris in August 2005.

In the rest of the paper, we talk about distributed biometric verification architectures (Section 2), summarize industry standards that are applicable to our framework (Section 3), describe the design and implementation of the framework (Section 4), show the principles of integration of external code into the MBAF (Section 5), and finally draw some conclusions and outline future work in Section 6.

## 2. DISTRIBUTED BIOMETRIC AUTHENTICATION

Many architectures exist for biometric verification systems. In abstract, functional terms, they mostly differ by how the processing steps for biometric data are divided between different machines. Here, we present a succint description of processing steps which are common to many biometric modalities.

**Capture or acquisition** The biometric data (voice, on-line signature, fingerprint, ...), also called biometric presentation, is digitised via the input device (microphone, pen tablet, fingerprint scanner, ...) and stored in memory.

**Preprocessing** The signal-domain acquired data is prepared for feature extraction. This is typically used for normalising the signal-domain data and remove biases or sources of corruption in a systematic fashion.

**Feature extraction** Discriminative features are extracted from the preprocessed data. Although features are very different for each biometric modality, the general underlying principle remains the same: this processing step typically reduces the dimensionality of the input data to create a feature-level representation of input patterns that will be used by the classifier to perform pattern recognition.

**Postprocessing** Features are normalised to remove bias or adapt them to the classifier.

**Template creation** User models, also called templates, are created from training feature sets to obtain a generic representation of a user that will be used for future comparisons. Many algorithms and can be used depending on

modality/features and model class, with some "models" corresponding directly to features.

**Background model creation** A background model, also called world model or anti-model, is needed by some biometric algorithms to provide normalisation for user presentation scores. They represent an "average" of the users from the population of the system.

**Template storage** Once their parameters are estimated, user models are stored in a secure location for use in later biometric operations.

**Template matching** A biometric presentation is compared with a particular user's biometric template. This typically results in a presentation score which is somehow related to how likely it is that the particular user is the source of that presentation. Depending on model and classifier types, this processing step will vary.

**Threshold computation** Several presentations belonging to a particular user and several presentations not belonging to that particular user (impostor presentations) are matched to that user's model to determine a hard limit (the threshold) below which a presentation will not be considered as belonging to the user.

The processing steps described above will be used in the following higher-level biometric *operations*, represented as an activity diagram in Figure 1.

**Enrolment** A user is added to the biometric system.

**Verification** The claim to a user's identity causes the presented biometric data to be compared against the claimed user's model.

**Identification** A database of user models is searched for the N most likely sources (N-best list) of the biometric presentation.

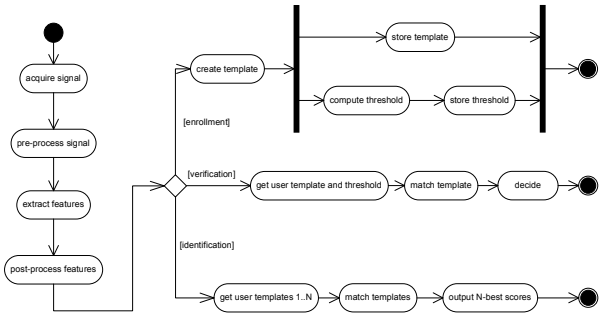



Figure 1: *biometric operations and corresponding processing steps for a generic modality*

Biometric system architectures mostly differ by how the steps (and hence operations) above are divided between different machines. In *off-line* systems, the access point has enough computing power and is trusted sufficiently to perform all the steps above locally. In *on-line* systems, the acquisition, preprocessing, and feature extraction are typically done on one of numerous client access points, and template creation, storage, and verification is performed on a biometric verification server.

Our framework defers all front-end tasks to the client, thus meaning that all pre-processing, feature extraction, and post-processing of feature is done locally and that only features travel through the wire. This offers significant advantages in terms of network bandwidth usage reduction and associated reduction of encryption overhead (see Section 2.3). It is however assumed that the client will have sufficient processing power to run a Java virtual machine and perform the needed operations, and have properly configured acquisition devices (microphone, pen tablet, fingerprint scanner...).

### 2.1. Middleware for distributed biometric applications

In previous work, we have higlighted that packet loss on IP networks would be detrimental for biometric verification over the internet, and advocated against the use of UDP for biometric data transmission. Furthermore, we have pointed out that packet reordering, a prevalent issue on the Internet, could also be damaging depending on the model and hence recommended using TCP [8].

Here we push this idea further by advocating for the use of middleware in our distributed biometric verification system, rather than doing a grounds-up implementation using socket programming and reinventing the wheel.

Object-oriented middleware allows objects on one host in a network to invoke a method on an object located on a different host. Typically, the middleware will handle at least data marshalling and unmarshalling, data types heterogeneity, and remote/local object references. Many choices are suitable for a distributed biometric verification system, for instance CORBA, COM, or Java RMI.

### 2.2. Java as middleware in biometric systems

In the presented framework we use Java RMI because it offers platform independence for both client and server objects and is less complex than CORBA to implement and deploy. Furthermore, the Java API offers many useful functionalities such as sound, cryptography and security functions and database access out of the box. Java is also flexible in offering binding mechanisms for native methods (the Java Native Interface or JNI) written in other languages, which means some parts of the system can be implemented in fast C++ code.

Lastly, if a three-tier architecture became necessary, for instance in the case of e-commerce transactions protected by biometric authentication, Java would offer an easy transition in the form of a Java Server Pages, applets, and web services combination.

### 2.3. Distributed verification security

Distributed biometric verification systems must contend with many challenges if they are to be made secure. Many of these challenges are shared by other distributed applications not related to biometric verification, and there is a large body of knowledge available to draw from in different fields [9].

In essence, biometric and application data should not be acquired or tampered with by third parties while being transmitted over the network (*leakage*), only authenticated clients should connect to authenticated servers, the systems should be able to cope with denial of service attacks, etc. Of critial importance is also the safeguarding of biometric templates and enrollment/adaptation data stored in the biometric database.

In this respect, the safest possible architecture is one where biometric templates or enrollment data are not stored in a central location, but instead stored on a smartcard-type device that each user carries around. The more biometric data a central

repository holds, the more its value will make fraudulent acquisition or tampering likely. The current system has no support for smartcard (distributed) template storage.

Java supports certificates, symmetric and asymmetric encryption, secure socket layer (SSL) connections and more security features through the Java Authentication and Authorization Service (JAAS) framework, the Java Cryptography Architecture framework, the Java Cryptography Extension (JCE), and the Java Secure Socket Extension (JSSE). Our framework uses both client and server X.509 certificates to authenticate both ends of the transaction. It also encrypts all communications between the client and server by using an SSL tunnel. Lastly, all biometric data stored on the database is encrypted using a symmetric encryption scheme, and decrypted on-the-fly as needed. We also use JAAS to implement user roles, so that administrators can enroll new users, but users can only verify themselves.

## 3. APPLICABLE INDUSTRY STANDARDS

### 3.1. The BioAPI specifiction

The BioAPI specification 1.1 (ANSI/INCITS 358-2002 standard) specifies data types, module functionalties, function names, data storage specifications and other items to promote interoperability of biometric technologies from different vendors and with different modalities. It defines high-level abstractions such as `Enroll` or `Verify`, which in turn can rely on primitive functions such as `Capture`, `Process`, `Match`, and `CreateTemplate`; these are assigned specific signatures and functionalities. The specification allows for much flexibility in the distribution of processing steps, and supports both off-line and on-line systems.

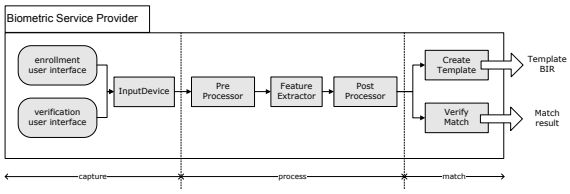



Figure 2: *Biometric Service Provider implementation*

One of the main concepts in BioAPI 1.1 is the Biometric Service Provider (BSP). As shown in Fig. 2, a BSP is responsible, for a given modality, to perform all the steps described in Section 2. Essentially, the BioAPI specification provides details of hooks into, and output data from, BSPs.

More recently, a new BioAPI specification (BioAPI v2.0) has been published [6]. It extends and generalises the BioAPI 1.1 framework. An interesting change is that it breaks down BSPs into sub-units called Biometric Function Providers (BFPs). To these functional level correspond "unit" levels, which replace the concept of "device" found in BioAPI 1.1. Thus, it is now possible for a BioAPI 2.0 BSP to use several sensors (sensor units).

The presented system uses an architecture that is inspired by the BioAPI specification in terms of object design and functional units, but is implemented in Java, and has no registered format owner for the biometric data. Thus, it is not BioAPI compliant. If needed, the system could be modified to make use of BioAPI-compliant BSPs by using the Java-to-BioAPI BSP bridge provided by Gens Software.

### 3.2. The ANSI X9.84 standard

The ANSI X9.84 standard [7], *Biometric Information Management and Security for the Financial Services Industry*, describes the security features needed to implement biometric verification for financial services. The Core Security Requirements are summarised here:

1. The integrity of biometric data and verification results must be guaranteed between any two components using software techniques such as hashes and physical measures such as tamper-resistent assemblies.
2. The source and receiver of biometric data and verification results must be authenticated, again using software or physical methods where appropriate.
3. The confidentiality of biometric data may be ensured between any 2 components.

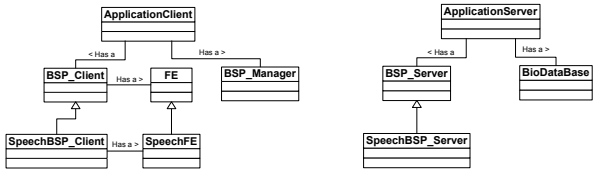
X9.84 also describes secure enrollment, verification, storage, transmission, and termination procedures.

Assuming the physical security of the machines our software, the security features exposed in section 2.3 make our framework compliant with the Core Security Requirements of X9.84.

## 4. FRAMEWORK DESIGN AND IMPLEMENTATION

### 4.1. Object hierarchy

The partial class diagram of Fig. 3 presents a simplified overview of the MBAF, showing only class derivations for the speech modality and hiding member data and methods.




(a) Client-side objects (b) Server-side objects

Figure 3: Partial class diagram with client-side and server-side components

The application uses three main objects: `BSP_Client`, which is intended to run on the access points such as desktop PCs, laptops, PDAs, ATMs, or other platform with IP network connectivity, Java support, and sufficient processing power. It is responsible for data acquisition, preprocessing, feature extraction, and user interface management. `BSP_Server`, which runs on the biometric verification server, typically a server PC running some flavour of Windows or Linux. This object is responsible for creating user templates, storing them in the biometric database, and scoring biometric presentations with respect to a particular user's template. `BSP_Manager` is responsible for loading and unloading of BSPs (modalities).

The `BSP_Client` is a base class that has one main class to perform front-end duties, `FE`. The latter has five generic classes, each responsible for an aspect of biometric data processing, matching the generic steps presented in Section 2. These are `InputDevice`, which can be specialised depending on the acquisition hardware (for the speech modality, `Microphone` is derived from `InputDevice`, and allows to set sampling rate, quantisation, etc.), `Preprocessor`, which can be specialised according to the modality and device being used,

`FeatureExtractor`, and `Postprocessor`. Each new modality added to the system can be implemented in Java using classes derived from the five generic classes above.

Similarly, on the server side, `SpeechBSP_Server` is derived from the `BSP_Server` base class because the `BioSPI_VerifyMatch` and `BioSPI_CreateTemplate` methods have some specificity for each modality, though in the case of speech or signature modelling the same model training engine (expectation-maximisation for learning Gaussian mixtures) could be used.

This object design is sufficiently generic to meet the needs of many other verification modalities such as fingerprints or face.

### 4.2. Biometric data storage

Biometric templates need to be stored in a database for the `BioSPI_VerifyMatch` to function. The framework provides format-agnostic features and model storage, retrieval, and deletion services. This is accomplished using the *MySQL* open-source database, to which the Java code connects using an open-source type IV Java Database Connectivity driver, `Connector-J`.

## 5. EXTERNAL MODULES AND NATIVE CODE

The framework uses the Java Native Interface (JNI) to be able to call methods on objects implemented in C++. The way external code is currently integrated into our framework is that JNI method signatures with JNI datatypes are added at global scope in the C/C++ source code of the biometric engine, the engine is recompiled into a shared object under Linux or a dynamic link library under Windows, and a minimal wrapper is written in Java to call on the native library. Biometric features and models are exchanged between the Java code and the native code by using the filesystem available on the server. While this may not be the most elegant solution, it offers the most flexibility for adding new biometric modules to our framework.

Work is currently underway to provide more options to integrate native code at different levels, for instance it might be interesting to use a native feature extractor somewhere in the processing chain. Also, a command-line wrapper structure will be added for the case where source code is not available.

### 5.1. Native library: Speech modality

The Torch C++ library [10] is used for training Gaussian Mixture speaker models and scoring speech presentations, and a JNI interface is used to bind to the `SpeechBSP_Server` Java object. File lists and feature files are simply put on disk and fed to the Torch engine.

On the client side, speech preprocessing consists of DC bias removal and silence removal. The features extracted are 12 MFCCs, deltas and delta-deltas. On the server side, speakers are modelled using a 64 components Gaussian Mixture models with diagonal covariance matrices World model normalisation is used.

### 5.2. Native library: On-line signature modality

The ALIZE C++ toolkit [3] is used to for training GMMs for signature modality users and scoring signature presentations. A similar JNI interface to that used for Torch links the native shared library to the Java framework.

The signature verification process is similar to that described in [11].

## 6. CONCLUSIONS

We have presented the design and implementation of a distributed framework for multimodal biometric authentication. The framework offers network transparency, end-to-end encryption, and biometric data management services as well as a skeleton for biometric authentication applications in the form of a set of base classes in Java which can be derived and extended.

Future work include making the integration of native code more flexible and with as little rewrite as possible (possibly none), making parallel modality use possible, and translating and improving the documentation and user manual.

The latest version of the MBAF with at least example speech and signature modalities can be downloaded from `http://scgwww.epfl.ch/jonas/software.html`.

## 7. ACKNOWLEDGEMENT

Part of this work was funded by a European Commission grant as part of the BioSecure network of excellence.

## 8. REFERENCES

[1] D.-Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, "SVC2004: First international signature verification competition," in *Proceedings 2004 Biometric Authentication: First International Conference, (ICBA 2004)*, (Hong Kong, China), pp. 16–22, July 2004.

[2] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book," tech. rep., Cambridge University Engineering Department. Available at http://htk.eng.cam.ac.uk.

[3] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, (Philadelphia, USA), pp. 737–740, March 2005.

[4] I. Nabney, *Netlab: Algorithms for pattern recognition*. Advances in pattern recognition, Springer, 2004.

[5] American National Standards Institute, *Information technology - BioAPI Specification (Version 1.1) (formerly ANSI INCITS 358-2002)*. New York, USA: American National Standards Institute, 2002.

[6] International Standards Organisation, *ISO/IEC 19794-1:2005 (BioAPI 2.0)*. Geneva, Switzerland: International Standards Organisation, 2005.

[7] American National Standards Institute, *Biometric Information Management and Security for the Financial Services Industry: ANSI X9.84-2003*. New York, USA: American National Standards Institute, 2003.

[8] J. Richiardi, J. Fierrez-Aguilar, J. Ortega-Garcia, and A. Drygajlo, "On-line signature verification resilience to packet loss in ip networks," in *Proc. 2nd COST 275 Workshop on Biometrics on the Internet: fundamentals, advances and applications*, (Vigo, Spain), pp. 9–14, March 2004.

[9] R. Anderson, *Security engineering: a guide to building dependable distributed systems*. John Wiley and Sons Inc., 2001.

[10] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[11] J. Richiardi and A. Drygajlo, "Gaussian mixture models for on-line signature verification," in *Proc. ACM Multimedia 2003 Workshop on Biometric Methods and Applications*, (Berkeley, USA), Nov. 2003.

# NON-INTRUSIVE FACE VERIFICATION BY A VIRTUAL MIRROR INTERFACE USING FRACTAL CODES *

*Ben A.M. Schouten, Johan W.H. Tangelder*
Centre for Mathematics and Computer Science (CWI), Amsterdam, the Netherlands
{B.A.M.Schouten, J.W.H.Tangelder}@cwi.nl

## ABSTRACT

Key factors in the public acceptance of biometric systems are non-intrusiveness, ease of use, and trust. In this paper we propose a biometric identity verification system consisting of a non-intrusive virtual mirror interface, and a face verifier using fractal coding to store the biometric template on a smart-card. The virtual mirror interface assists the user to obtain a frontal face image. The limited memory requirements of the fractal template and processing requirements of fractal decoding enable the implementation of the verification procedure on a smart-card. This set-up facilitates non-intrusive and easy to use biometric verification, and provides trust by adding a visualization to the biometric yes/no decision. Since the system does not require user assistance, it enables secure access over the Internet.

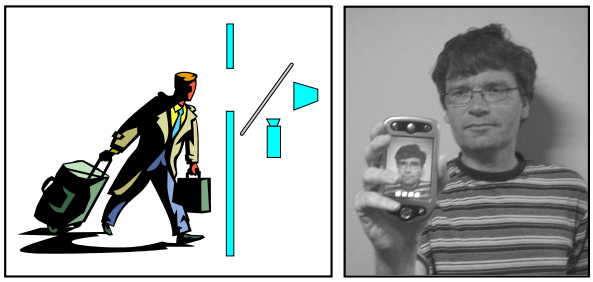


**Figure 1:** Supporting biometrics over the Internet with a virtual mirror interface: (a) from a public access point, and (b) from a mobile phone. The webcam and video display share the same optical axis through a half-silvered mirror.

## 1. INTRODUCTION

User convenience, non-intrusiveness, and trust are key factors for the public acceptance of biometric systems [14]. Therefore, authentication should be as easy as possible for the user, and the user should be able to monitor the authentication process. In this paper we present a smart mirror interface meeting these requirements. The system can be applied to provide secure access over the Internet from home, from mobile phones, or from public access points to personal data of the user, see figure 1. Moreover, the limited memory and processing requirements, facilitate the application of the system in advanced mobile phones.

The virtual mirror interface, which consists of a display visible through a half-silvered mirror, supports all user interaction in a natural way. The half-silvered mirror mirrors the user's image both back to the user, and to a webcam. To enroll the user has only to look into the mirror, enabling the detection of a frontal face image by the webcam. Fractal encoding is applied to compute a compact biometric template of the face image. This biometric template of the user can for instance be stored on a smart-card. To verify his/her identity the user presents the smart-card and looks again in the mirror and only has to align his face with a mask (with the same pose as the enrollment image) shown on the display behind the mirror, see figure 2. Again, a frontal face image is detected by the webcam. Next, the face verifier applies fractal decoding to morph on the display the new detected face to the enrolled face, see figure 3.

An advantage of using fractal coding of the enrolled face image is the small size of the template which fits on a smart-card. Moreover, morphing the actual presented image to the previous enrolled image of the user, enhances trust in the biometric verification process.

In the next section we describe the virtual mirror interface and a user scenario. In section 3 we discuss fractal coding and in section 4 we present experimental results on using fractal codes for face recognition. We conclude and indicate directions for further research in section 5.

## 2. INTERFACE AND USER SCENARIO

Darrell et al. [3] demonstrate the application of a virtual mirror interface in an interactive entertainment system displaying

*Part of this research has been funded by the Dutch BSIK/BRICKS project.

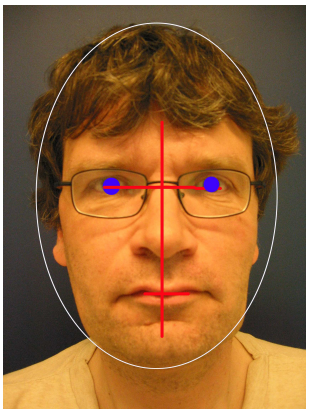


**Figure 2:** The user is invited to present his face to the system, guided by a mask denoting a frontal pose.

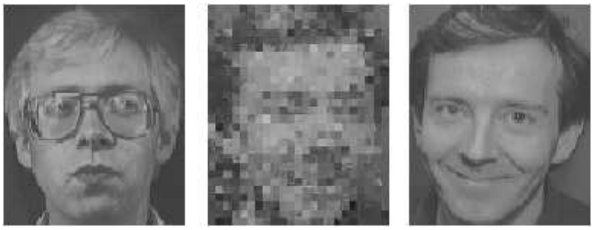


**Figure 3:** The face image of an impostor is morphed into the face of the authorized user of an application, showing three intermediate steps.

a user's face distorted by various graphical effects. We propose to use the virtual mirror interface to provide secure access over the Internet and guide the user in the authentication process. Figure 1 illustrates our set-up. The virtual mirror is created by placing a camera sharing the same optical axis as the video display, using a half-silvered mirror to merge the two optical paths. The camera views the user through a right-angle half-silvered mirror, so that the user can view a monitor while also looking straight into (but not seeing the camera).

At enrollment the user is invited to present his face in front of the mirror. A face extractor detects and scales the face of the user. To obtain the face in a frontal pose, the user has to align his face with a mask as illustrated by figure 2. On the display the mask is superimposed on the user's face. If the face and the mask align, a biometric template is extracted from the face, e.g. a fractal code of the presented face as explained in the next section. The template can be stored on a smart-card and in that case the system prints a smart-card containing the template.

For verification the client or the impostor presents her/his face and (stolen) smart-card to the smart mirror of the same or another intelligent kiosk or device. To visualize the verification process a morph from the presented face to the true face is shown as illustrated by figure 3.

# 3. FACE RECOGNITION USING FRACTAL CODES

Several face recognition techniques, such as principal components analysis, active appearance models, or elastic bunch graph matching, can be applied for biometric verification. Unfortunately, due to memory limitations these methods are sometimes difficult to implement on a smart-card. Therefore, to the problem of face recognition we would like to apply fractal coding, which is a relatively new technique which emerged from fractal geometry [7]. It is recognized that the search for similarities is best approached by a fractal mechanism and can be used to create fractal features, invariant under scaling, small rotations and translations [2, 4, 5, 6, 12]. An advantage of using fractal coding is the small size of the fractal code, which fits on a smart-card and the possibility of on-card processing.

Fractal coding [2, 4] is based on the self-similarity in a picture, which means that (small) regions of a picture can be transformed versions of some other (larger) regions of the same picture. *Partitioned iterated function systems* (PIFSs) encode this self-similarity by a limited set of affine transformations on the support and gray values of the image. The number of functions in the system is large, typically hundreds. In this paper we examine the properties of these fractal functions for the use of face recognition.

At encoding, an image $f$ is encoded by finding a transformation $W$ and an image $\hat{f} \approx f$ for which

$$W(\hat{f}) = \hat{f}. \tag{1}$$

In order to do so, a given image is partitioned into non-overlapping range blocks. The fractal encoder searches for parts called domain-blocks (which can be larger and overlapping) in the same image that looks similar under some fixed number of affine transformations. Such an affine transformation can be written as:

$$w_i(\vec{x}) = A_i\vec{x} + \vec{o}, \qquad A_i \equiv \begin{pmatrix} a_i & b_i & 0 \\ c_i & d_i & 0 \\ 0 & 0 & u_i \end{pmatrix}, \tag{2a}$$

$$\vec{x} \equiv \begin{pmatrix} x \\ y \\ f(x,y) \end{pmatrix}, \quad \vec{o} \equiv \begin{pmatrix} s_i \\ t_i \\ o_i \end{pmatrix}, \|u_i\| \leq 1. \tag{2b}$$

Index $i$ indicates the range-blocks within the image, $f(x,y)$ denotes the gray-value at position $(x,y)$. $u_i$ is a contrast scaling on the gray-values and $o_i$ is a luminance offset on the same

values. The parameters $a_i, b_i, c_i, d_i, s_i, t_i$ constitute a geometric transform on the support of the image. The parameters $u_i$ and $o_i$ are used to match the gray-values of the domain with the gray-values of the range-block, within the limits of an imposed accuracy ε. Usually, a domain-block has twice the size of a range-block. The contractive nature of the transformations $w_i$ makes the fractal encoder work. In the encoding all parameters describing the transformation $W = \bigcup_{i=1}^{N} w_i$ (where $N$ is the total number of range blocks in the image) are stored.

According to the Contractive Mapping Fixed-Point Theorem [4], the fixed point $\hat{f}$ of $W$ can be restored by iterating $W$ in the decoding phase starting with an arbitrary given image $f_x$; with every iteration new detail is created.

$$\hat{f} \equiv \lim_{n\to\infty} W^{\circ n}(f_x);\ W^{\circ n}(f_x) = \underbrace{W \circ \cdots \circ W(f_x)}_{n \text{ times}} \qquad (3)$$

Fractal theory has found many applications in image processing. Particularly, PIFSs have received a great deal of attention in image compression [2, 4, 5] and fractals have been applied to object and/or face recognition [6, 12]. Two basic approaches are found in the literature. In the first approach [1, 8, 9, 11, 13], features extracted from the fractal code of the input object are compared against features extracted from the fractal codes of the objects stored in the database. The second approach [6, 12], is based on comparing distances, using the Euclidean distance measure, between one object and another object modified by one iteration using a fractal code.

Both methods, may be applied for biometric identification. However, only the latter method allows to morph the face presented to the virtual mirror to the genuine face. Therefore, in the remainder of the paper we focus on a method modifying the input object using the fractal code derived from the face image of the user $g$, whose identity is to be verified.

### 3.1. Fractal Feature Extraction for Virtual Mirror Interfaces

At enrollment a picture $f_g$ is taken from the client $g$. A fractal code $W_g$ is generated from $f_g$ for which

$$W(f_g) \approx f_g. \qquad (4)$$

In the first state of the authentication process, the user $i$, not necessarily being the genuine user, claims the identity $g$ and presents a (stolen) smart-card to the reader. The system reads the fractal codes $W_g$ from the smart-card and generates the previous enrolled image of the client, or in fact the fractal estimation of the enrolled image, $\hat{f}_g$.

However, this image is not presented to the user, as this would immediately unmask a possible impostor before the authentication process is completed. Instead a mask as shown in figure 2, is used to guide the user to the authentication process and to guarantee that a picture $f_i$ can be taken in the same pose as the enrolled image, improving the robustness of the application.

During the authentication process we calculate:

$$d(W_g(f_i), \hat{f}_g) \qquad (5)$$

where

$$d(f,g) = \sqrt{\sum_{k=0}^{I_h}\sum_{l=0}^{I_w}(f^{k,l} - g^{k,l})^2}, \qquad (6)$$

and $f^{k,l}$ denotes the pixel value of image $f$ at position $(k,l)$. $I_h$ and $I_w$ are the height and width of the image. According to the threshold set in the system, the identity of the user is verified.

Now, in the next step, as a result of the authentication process, an image $W_g^{\circ 10}(f_i)$, which resembles the enrolled image sufficiently, is generated and presented to the user and/or supervisor of the system, showing the similarity between the enrolled image and the presented image and allowing the user to see for him self if the authentication process was successful.

All intermediate steps are shown resulting in a morph from the enrolled image to the presented image at the time of the authentication process, as illustrated by figure 3. In general ten iterations is enough to render a clear approximation of the previous enrolled image.

## 4. EXPERIMENTAL RESULTS

The aim of this section is to investigate the feasibility of the fractal coding for biometric authentication in a virtual mirror interface. For our experiments we used the fractal coding software, which is described in the book by Fisher [4]. The Olivetti Research Labs (ORL) face database [10] was used in our experiments. Although, the ORL database is considered as an easy database, it was used because the faces are centred on the image, so face detection was not required, and it contained faces of varying pose and expressions. The ORL face database contains ten different images of each of the 40 distinct subjects. To investigate the robustness of the fractal distance measure with regard to pose, we created two databases containing images with similar pose. For the first database we selected per subject 2 images with similar pose, and for the second database we selected per subject 3 images.

For all images in the complete database, we used the other 9 images of the subject in the database to evaluate client claims and the other 390 images not of the subject to evaluate impostor claims. For the complete database this evaluation procedure resulted in 3600 client claims and 156.000 impostor claim.

For all images in the two pose similar databases, we used the other image (two images) of the subject to evaluate client claims and the other 78 images (117 images) to evaluate impostor claims. For the first pose similar database this evaluation procedure resulted in 80 client claims and 6.240 impostor claims, and for the second pose similar database in 240 client claims and 14.040 impostor claims.

We plotted for the three databases the receiver operating characteristic (ROC) curves in figure 4. The ROC curves show that using pose similar images, improves performance: from an EER of 12.3 % for the ORL database, to an EER of 6.9 % for the subset containing three images per subject, and an EER of 6.0 % for the subset containing two images per subject.

However within our experiments we did not use the automatic pose correction. In future research we will test the results, generating a new database using the pose guidance system and expect the results to further improve. Also, instead of using the default parameter settings of the fractal coding software we would like to optimize its parameter setting to improve the results.

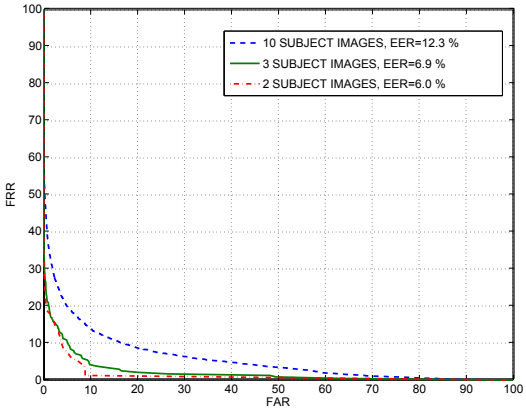



**Figure 4:** Plot comparing ROC curves using the full ORL database with 10 images per subject, a subset of 3 frontal pose images per subject, and a subset of 2 frontal pose images per subject.

# 5. CONCLUSIONS AND FURTHER RESEARCH

In this paper a virtual mirror interface supporting secure access over the Internet, is proposed. The virtual mirror interface assists the user to obtain a frontal face image. Because the system does not require user assistance, it facilitates secure access over the Internet. Since fractal coding provides compact codes, we apply face recognition based on fractal coding to implement the proposed scenario using smart-cards. Fractal methods could be used for on card processing which improves both privacy and security within the biometric authentication process. To evaluate the recognition performance for frontal poses we used the ORL database [10], and we selected two subsets from the ORL database, consisting of face images in a frontal pose. Our results show that the recognition performance of fractal coding improves using automatic pose reconstruction.

Fractal codes achieve very compact coding, compressing each image from the ORL face database from 10.1 K to a range of 1-1.8 K fractal code. Hence, fractal compression of features obtained by filtering images, for instance using Gabor filters, is a promising research direction to obtain both compact biometric templates and high recognition rates.

Also, to improve recognition rates, fusion of speaker and face modalities using the relatively low quality images and low quality sound from a webcam, is an important research issue. For liveness detection video can be applied to detect the presence or absence of spontaneous facial expressions changing the face. For high security applications, the system can ask the user to show facial expressions, e.g. blinking, which are difficult to spoof.

# 6. REFERENCES


[1] Baldoni, M., Baroglio, C., Cavagnino, D., Egidi, L.: *Learning to Classify Images by Means of Iterated Function Systems.* In: Fractals and Beyond: Complexities in the Sciences, 173-182, World Scientifi c, 1998.

[2] Barnsley, M.F., Hurd, L.P.: *Fractal Image Compression.* AK Peters Ltd., 1993

[3] Darrell, T., Gordon, G., Woodfi ll J., Harville, M.: *A Virtual Mirror Interface using Real-time Robust Face Tracking.* In: Proc. of the Third International Conference on Face and Gesture Recognition, April 14-16, 1998, Nara, Japan.

[4] Fisher, Y. (ed.): *Fractal Image Compression, Theory and Application.* Springer Verlag, 1995.

[5] Jacquin, A.E.: *Fractal image coding: a review.* in: Proc. of the IEEE, 81 (10), 1451-1465, 1993.

[6] Kouzani, A.Z., He, F., Sammut, K.: *Towards Invariant Face Recognition.* In: Information Sciences 123, 75-101, 2000.

[7] Mandelbrot, B.B.: *The Fractal Geometry of Nature.* Freeman and Company, New York, 1983.

[8] Marie-Julie, J.M., and Essafi , H.: *Image Database Indexing and Retrieval Using the Fractal Transform.* In: Proc. of Multimedia Applications, Services and Techniques, 169-182, Springer Verlag 1997.

[9] Neil, G., Curtis, K.M.: *Scale and Rotationally Invariant Object Recognition using Fractal Transformations.* In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 3458-3461, 1996.

[10] Samaria, F.S., Harter, A.C.: *Parameterisation of a stochastic model for human face identi cation.* In: 2nd IEEE Workshop on Applications of Computer Vision, 1994

[11] Schouten, B.A.M., de Zeeuw, P.M.: *Fractal Transforms and Feature Invariance.* In: Proc. of the International Conference on Pattern Recognition (ICPR'00), 2000.

[12] Tan, T., Hong Y.: *The Fractal Neighbor Distance Measure.* In: Pattern Recognition 35, 1371-1387, 2002.

[13] Vissac, M., Dugelay, J., and Rose, K.: *A Fractals Inspired Approach to Content-Based Image Indexing.* In: Proc. IEEE International Conference on Image Processing, on CDROM, 1999.

[14] *Privacy best practices.* In: Biometric Technology Today, Vol. 11, Issue 11, 8-9, 2003.

# EVALUATING BIOMETRIC ENCRYPTION KEY GENERATION

*Michael Fairhurst, Sanaul Hoque, Gareth Howells, Farzin Deravi*

Department of Electronics, University of Kent, Canterbury, United Kingdom.
E-mail: {M.C.Fairhurst,S.Hoque,W.G.J.Howells,F.Deravi}@kent.ac.uk

## ABSTRACT

In traditional cryptosystems, user authentication is based on possession of secret keys/tokens. Such keys can be forgotten, lost, stolen, or may be illegally shared, but an ability to relate a cryptographic key to biometric data can enhance the trustworthiness of a system. This paper presents a new approach to generating encryption keys directly from biometrics.

## 1. INTRODUCTION

Internet transactions are commonplace in modern society and, increasingly, issues surrounding the security of such transactions are being seen as the major factors in promoting the effective use of networked systems. The ubiquity of the Internet and the vulnerability of digital data make it even more difficult to control and trace intrusions by unauthorised individuals, and it is therefore essential that e-commerce applications provide a secure mechanism for accurate user identification and controlled access to sensitive data.

Modern encryption techniques can, of course, help to address these issues and PKI incorporating passwords, PIN or smartcards are well established. However, the security of a robust cryptographic system lies in the ability to keep the cipher keys secret. Most users typically choose trivial or obvious passwords, and thus may compromise the security of the systems while, in any case, such techniques cannot identify a user beyond doubt. Therefore, a user authentication scheme, perhaps involving a biometrics-related technique, may usefully be incorporated.

It is therefore logical to investigate the possibility of the merger of the present two-tier approach into a one-step process in such a way that encryption key(s) are extracted directly from the biometric samples. Having a biometric-based key offers a number of distinct advantages compared to more traditional approaches, including the fact that this removes the need for a user to remember his keys and, as a biometric is an inherent feature of a human user, it cannot be transferred to a third party. However, such an approach also raises a particular difficulty, which is that conventional encryption algorithms clearly require the cipher-key to be exactly the same on every instance of use. The nature of biometric data is such that this requirement is generally difficult or impossible to satisfy in practice. In addition, factors such as ageing, illness, environmental conditions, and other influences have a bearing on the quality and variability of typical biometric data, and thus direct repeatability is a major issue in this type of situation.

In this paper we discuss a possible scheme for generating encryption keys directly from biometric samples which takes into account the practical issues noted.

## 2. LITERATURE REVIEW

The notion of using biometrics directly as a cryptographic key has been of interest for a number of years now and a number of approaches have been proposed in the literature. Uludag *et al* [1] recently published a detailed discussion of the issues and challenges of biometric cryptosystems. The principal obstacle for such a system is the inherent variability of user biometrics and to overcome this, the most common approach is to bind the keys with the biometric templates as payloads. A positive biometric identity verification transparently releases this key. Many commercial biometric programs support this approach, and Soutar *et al* [2] also introduced a similar scheme. Monrose *et al* [3] suggested a method of securing passwords by combining with keystroke dynamics (named *hardened password*). An encrypted instruction table along with a history file is incorporated to validate the authenticity of the user and create the hardened password to be used as the key. They also suggested a similar but more refined approach using voice biometrics [4], and another approach would be to generate the key from this template using a suitable one-way function. Bodo [5] first proposed that data derived from the template be used directly as a cryptographic key and Janbandhu *et al*. [6] and Daugman [7] supported this notion. As sample variability has no direct bearing on these templates, the same key can be generated at all times, but a major drawback of the approach is that if a user needs to change his template (due to ageing, for example), the previous key may never be regenerated. Uludag *et al* [8] proposed a multi-phase protocol where a key is developed during the initial handshaking with the server. The key is a function of the data as well as the server and user identity (biometric). This scheme assumes the presence of a trusted handler that prevents decryption if biometric verification failed. Vector quantization (VQ) can be used to counter the variations in a user's biometric samples. The feature space is partitioned into a number of cell spaces, each cell space denoted by mean vectors. Live samples are compared with these vectors. The nearest vector in the code-book determines the user identity and then can be used to extract the encryption key. Yamazaki *et al*. [9] described such a method to generate a biometric-based key.

## 3. THE PROPOSED SCHEME

Here we introduce a different approach. The method proposed creates a series of partitions in the feature subspaces, each represented by a subset of feature dimensions where the features are extracted from a biometric sample. The partitions define a number of cells in these subspaces and each cell is tagged with a key component (usually its identity). When a

live biometric sample is available, it is checked against these partitions to ascertain its membership of a cell. Each feature subspace (denoted by its own set of partitions) is treated independently and contributes a component in the encryption key. As there are generally many subspaces, by concatenating these individual key segments, a complete key can be obtained. In the proposed scheme, users do not need to declare their individual identities to have access to a secured file. The capability to provide a biometric sample which can decipher the file is regarded as acceptable evidence of proof of identity. On the other hand, it must also be recognized that the partitions are created based on feature-point distributions in the subspace (rather than user-identities), and therefore multiple users may, in principle, share the same cell-space. In such a situation their biometrics will lead to the same encryption key. However, such unintended "impersonation", where an individual tries to access a document secured by a cell-mate, is found very unlikely [10].

## 3.1 Complexity

Let $X_i^B$ denote an $n$-dimensional feature subspace extracted from an $N$-dimensional feature vector $\overline{X}^B$ obtained from a given biometric $B$, and

$$X_i^B = \{x_i^1, x_i^2, \ldots, x_i^n\}.$$

Let this subspace be partitioned into $k_i$ cells.

This feature subspace can generate a key component $\lambda_i$ of $(\log_2 k_i)$-bits. By concatenating all the key components from all available subspaces (or a subset of these), we are in a position to generate the key

$$K = \text{concat}(\lambda_i), i = 1, \ldots, l.$$

The key size will be $\sum \log_2 k_i$, $i = 1, \ldots, l$ where '$l$' is the number of possible ways of creating feature subspaces. For $\overline{X}^B$,

$$l = {}^N C_1 + \ldots + {}^N C_N = \sum_{i=1}^{N} {}^N C_i, \text{ where } {}^n C_r = \frac{n!}{r!(n-r)!}.$$

When $N = 5$, $l = 31$. For $k_i = 8$, we have 3-bit key string for each subspace and the total key-size is 93 bits. By adjusting $N$ and $k_i$ judiciously, we can easily modify the overall key size.

In addition, calculation of the key requires only $l(k+1)$ comparisons of $n$-dimensional vectors. For $N = 5$ and $k_i = 8$, a 93-bit key can be obtained from 279 operations.

# 4. EXPERIMENTS AND RESULTS

The empirical investigations reported here have been designed to achieve two broad objectives. Firstly, we wish to establish the viability of the proposed technique using synthetic feature vectors and also to determine the essential feature attributes for such a system. Subsequently we wished to carry out investigations to determine performance when real, rather than synthetic, biometric data is used.

The synthetic database consisted of statistically-modelled feature vectors for 400 users each giving 10 samples. The user data are assumed uniformly distributed in the feature space with each individual showing Gaussian intra-class distribution. The standard deviation associated with individual user data generated at random but restricted to a pre-determined maximum. A randomly selected set of 5 samples from each user were used for training and the remainder for testing. This synthetic database, therefore, represents a near-idealistic scenario.

For a real biometric database, handwritten signatures have been used. The experimental data were collected during the public trials of automatic signature verification undertaken by the University of Kent in association with the British Technology Group. The signatures were collected from a cross-section of the general public in a typical retail environment. Signature samples were collected over multiple sessions. The overall database consists of 7430 signatures contributed by 359 volunteers though, for practical reasons the number of signatures donated at each trial was not fixed. In order to ensure sufficient training data, a subset of 273 users were selected, comprising those volunteers who contributed at least 10 samples each.

The advantage of using the handwritten signature as an example of a real biometric is that it corresponds to a modality which has been in use widely for many years and enjoys maximum user-acceptability. The signature also exhibits a major disadvantage, however, in that, being a behavioural biometric, it typically shows very high intra-class variations. Thus, the experimental database adopted may be considered to be representative of a near-worst case scenario within which the proposed approach can be evaluated.

The proposed approach was first applied to the synthetic data, and Table 1 demonstrates the extent to which a genuine user fails to be positioned in the cell-space assigned to him (corresponding effectively to a false rejection rate, FRR, for the system). This cell assignment was determined during training as the subspace cell containing most of his training samples. It is clear from Table 1 that, on average, the FRRs generated more than doubled as the number of partitions was doubled. At the same time, the system FRRs also deteriorated by more than a factor of 2 when $\sigma_u$ (intra-user standard deviation) is doubled. Hence it is clear that, when intra-user variability is high, it is better to use fewer partitions, and variability beyond a certain limit renders that feature subspace useless because the associated errors accumulate when key segments from other subspaces are concatenated. Ideally, in order to create the conditions which will lead to an acceptable system performance in terms of the FRR generated for 1-D subspaces, it is clear that we should aim to satisfy the condition $\sigma_u << 1/k_i$.

| Max($\sigma_u$) | No of cell partitions, $k_i$ | | | |
|---|---|---|---|---|
| | 2 | 4 | 8 | 16 |
| ≤ 0.005 | 0.12 | 0.39 | 1.01 | 2.31 |
| ≤ 0.01 | 0.39 | 1.15 | 2.76 | 6.28 |
| ≤ 0.02 | 0.98 | 2.32 | 5.33 | 12.59 |
| ≤ 0.05 | 2.27 | 6.56 | 17.48 | 31.51 |
| ≤ 0.10 | 4.56 | 13.96 | 32.08 | 48.51 |

**Table 1:** Mean FRR (%) from partitioning 1-D feature vectors (for synthetic data).

We also investigated the effect on FRR of concatenation of multiple subspace identities, as shown in Table 2. Similar key-spaces can be generated using different settings, resulting in significantly different FRR. For example, a 32 bit key can be generated either from $\{N = 32, k_i = 2\}$, $\{N = 16, k_i = 4\}$ or $\{N = 8, k_i = 16\}$ resulting in FRRs of 12.6, 16.5 and 41.1 respectively (for $\sigma_u$ = 0.01). The same pattern is observed for other $\sigma_u$, emphasising the importance of a large $N$. Since most robust applications require much larger keys, Table 3 shows performance when, for example, a 128-bit key is required.

| max $\sigma_u$ | $K_i$ | FRR | | |
|---|---|---|---|---|
| | | ***N*=8** | ***N*=16** | ***N*=32** |
| 0.005 | 2 | 0.9 | 2.2 | 5.6 |
| | 4 | 3.1 | 5.3 | 13.7 |
| | 8 | 7.8 | 16.8 | 27.1 |
| | 16 | 17.9 | 30.8 | 51.0 |
| 0.01 | 2 | 3.2 | 6.3 | 12.6 |
| | 4 | 8.8 | 16.5 | 27.3 |
| | 8 | 20.0 | 35.8 | 54.0 |
| | 16 | 41.1 | 63.7 | 83.1 |

**Table 2:** Mean FRR (in %) after concatenation of 1-D partitions

| Configuration | FRR | | | |
|---|---|---|---|---|
| | **$\sigma_u$ = 0.001** | **$\sigma_u$ = 0.005** | **$\sigma_u$ = 0.01** | **$\sigma_u$ = 0.02** |
| {*N*=128, $k_i$=2} | 1.3 | 19.5 | 37.5 | 65.8 |
| {*N*=64, $k_i$=4} | 3.3 | 25.0 | 50.6 | 76.1 |

**Table 3:** Mean FRR(%) for a 128-bit key

It is apparent, therefore, that useful encryption keys may be generated from biometric measurements depending on the characteristics of the extracted features. In order to demonstrate the application to real data, features extracted from handwritten signatures were next considered. Table 4 shows the complete feature-set used in this study.

The 'Mean' and 'std' of user means show the principal concentration and spread of signers in feature space. Ideally, users should be uniformly distributed over the whole of the available feature space. This would give a mean near to 0.5 and very large std. This can be maintained in the synthetically generated data, but very few real features (e.g., features 14-16, 24 etc) achieved this. However, their low std indicates that there is more of a concentration near the centre rather than uniform distribution over the feature space. The other real data also revealed similar trends, but with the centre of mass at different locations. The implication of these observations is that special measures might be helpful in counteracting the bias, although no such measures were incorporated for the results presented here. The last column emphasises the intra-class attributes. Both of these measures illustrate user data variability and the effect of this.

Despite the non-ideal distribution of the extracted features, the feasibility of creating useful 1-D partitions can be investigated. Table 5 shows the success rates (in %) in identifying the correct partition using the extracted features. The FRRs are rather high in most cases due to their high average of intra-class variations. However, some features (e.g., features 4, 12, 13, 19, etc) produced much lower FRR. Due to the small numbers of usable features, the effect of concatenation of the 1-D spaces was not further investigated at this stage.

| | **Feature description** | **Average user means, $\overline{\mu_u}$** | **St dev of user means,** | **Average of user std, $\overline{\sigma_u}$** |
|---|---|---|---|---|
| 1 | Width-height ratio | 0.197 | 0.108 | 0.031 |
| 2 | Total execution time, $T_s$ | 0.187 | 0.086 | 0.023 |
| 3 | No. of strokes | 0.216 | 0.123 | 0.035 |
| 4 | Mean horizontal velocity, $v_x$ | 0.104 | 0.098 | 0.013 |
| 5 | Mean vertical pen velocity, $v_y$ | 0.241 | 0.143 | 0.030 |
| 6 | Vertical mid-point crossing | 0.336 | 0.138 | 0.063 |
| 7 | Horizontal centralness, $m_{20}$ | 0.174 | 0.106 | 0.031 |
| 8 | Vertical centralness, $m_{02}$ | 0.116 | 0.102 | 0.029 |
| 9 | Diagonality, $m_{11}$ | 0.221 | 0.128 | 0.042 |
| 10 | Horizontal divergence, $m_{12}$ | 0.090 | 0.085 | 0.026 |
| 11 | Vertical divergence, $m_{21}$ | 0.114 | 0.088 | 0.028 |
| 12 | Horizontal imbalance, $m_{30}$ | 0.085 | 0.073 | 0.021 |
| 13 | Vertical imbalance, $m_{03}$ | 0.055 | 0.072 | 0.019 |
| 14 | $\frac{duration\ of\ v_y > 0}{T_w}$ (pen down only) | 0.536 | 0.113 | 0.054 |
| 15 | $\frac{duration\ of\ v_y < 0}{T_w}$ (pen down only) | 0.524 | 0.157 | 0.051 |
| 16 | $\frac{duration\ of\ v_x > 0}{T_s - T_w}$ (pen up only) | 0.551 | 0.207 | 0.087 |
| 17 | $\frac{duration of\ v_x < 0}{T_s - T_w}$ (pen up only) | 0.270 | 0.171 | 0.085 |
| 18 | $\frac{duration of\ v_y > 0}{T_s - T_w}$ (pen up only) | 0.276 | 0.142 | 0.078 |
| 19 | $\frac{\overline{V}}{\max\lvert v_y\rvert}$ | 0.116 | 0.047 | 0.022 |
| 20 | $\frac{signature\ length}{signature\ area}$ | 0.183 | 0.097 | 0.037 |
| 21 | Total $\lvert v_x = 0\rvert$ event | 0.256 | 0.132 | 0.048 |
| 22 | Total $\lvert v_y = 0\rvert$ event | 0.247 | 0.111 | 0.042 |
| 23 | $\frac{duration\ of \min\lvert v_y\rvert}{T_w}$ (pen down only) | 0.347 | 0.144 | 0.058 |
| 24 | Initial direction | 0.584 | 0.141 | 0.181 |

**Table 4:** Characteristics of the features extracted from real signatures.

| Feature # | $k_i$=4 | $k_i$=8 |
|---|---|---|
| 1 | 10.15 | 23.97 |
| 2 | 4.26 | 13.92 |
| 3 | 11.3 | 21.8 |
| 4 | 1.43 | 6.56 |
| 5 | 9.82 | 22.11 |
| 6 | 24.2 | 39.32 |
| 7 | 7.63 | 20.1 |
| 8 | 6.32 | 17.6 |
| 9 | 14.47 | 30.79 |
| 10 | 3.65 | 13.45 |
| 11 | 4.98 | 16.02 |
| 12 | 2.64 | 10.31 |
| 13 | 2.15 | 9.3 |
| 14 | 19.4 | 39.76 |
| 15 | 18.47 | 36.95 |
| 16 | 25.87 | 46.01 |
| 17 | 22.61 | 41.2 |
| 18 | 22.58 | 44.38 |
| 19 | 0.92 | 22.51 |
| 20 | 10.54 | 26.73 |
| 21 | 16.44 | 31.68 |
| 22 | 14.68 | 32.3 |
| 23 | 19.45 | 39.5 |
| 24 | 52.45 | 73.16 |

**Table 5:** Mean FRR (%) from partitioning 1-D feature vectors (signature data).

## 5. DISCUSSION

This paper has introduced a method for generating encryption keys directly from live biometrics. Initial investigation using synthetic features revealed that, under appropriate conditions, it is possible quite accurately to extract a unique cipher key for use with standard encryption algorithms. The situation is, however, more complex with real biometrics. Behavioural biometrics (such as the signature), especially, show very high variation in the measured features, and thus the corresponding FRR, even when using a 1-D partition, are likely to be significant, as our worst case scenario demonstrated. Despite this, a number of features could be found which had relatively low variance indicating, and this raises the possibility that careful feature analysis and selection can lead to successful adoption of this approach.

It is also apparent that the capability of creating many partitions from each feature will enable generation of a large enough key from fewer features, although very low variance features are not common in practice, and effort should be directed towards minimal partitioning using many features. Using physiological biometrics may be a better alternative. Concatenation of key components extracted from multiple biometrics can also help in generating large keys.

A number of possibilities for improving performance are foreseen. One is to filter training samples to remove outliers before use for partition boundary detection. A region in the vicinity of the boundaries may be marked as an unsafe region, and when a measured feature falls into this area, additional samples need to be captured. Filtering out users with unstable biometrics (goats) can also improve the performance for other users.

## 6. REFERENCES


1. U. Uludag, S. Pankanti, S. Prabhakar and A.K. Jain. Biometric Cryptosystems: Issues and Challenges. *Proceedings of the IEEE*. **92**(6):948-960. 2004.

2. C. Soutar, *et al.* Biometric Encryption. In R.K. Nichols (ed.): *ICSA Guide to Cryptography*. McGraw-Hill. 1999.

3. F. Monrose *etal.* Password hardening based on keystroke dynamics. In *Proc. of ACM conf computer and communication security*. Singapore. Pages 73-82, November 1999.

4. F. Monrose *et al.* Cryptographic key generation from voice. In *Proc. 2001 IEEE Symposium on Security and Privacy*. Oakland, CA, USA. Pages 202-213. May 2001.

5. A. Bodo. *Method for Producing a Digital Signature with Aid of a Biometric Feature*. German Patent DE 4243908A1. 1994.

6. P. K. Janbandhu *et al*. Novel Biometric Digital Signatures for Internet-based Applications. Information Management and Computer Security. Vol. 9(5), pp.205-212. 2001.

7. J. Daugman. *Biometric Decision Landscapes*. Technical Report TR482. University of Cambridge Computer Laboratory, Cambridge, UK. 2000.

8. U. Uludag *et al*. Multimedia Content Protection via Biometrics-based Encryption. In *Proc. ICME2003*. Maryland, USA. 2003.

9. Y. Yamazaki *et al*. A Secure Communication System using Biometric Identity Verification. *IEICE Trans. Information And Systems*. Vol. **E84-D**(7), pp.879-884. 2001.

10. S. Hoque *et al.* On The Feasibility of Generating Biometric Encryption Keys. *IEE Electronics Letters*. **41**(6):309-311. 2005.

# SPEAKER SEGMENTATION, DETECTION AND TRACKING IN MULTI-SPEAKER LONG AUDIO RECORDINGS

*Laura Docío-Fernández and Carmen García-Mateo*

E.T.S.I. Telecomunicacion, Departamento de Teoria de la Señal y Comunicaciones, University of Vigo (Spain)
ldocio@gts.tsc.uvigo.es, carmen@gts.tsc.uvigo.es

## ABSTRACT

In this paper we present an approach for speaker identification or detection and tracking in a task where the audio stream contains speech of several speakers. The approach is based on a system that performs an audio segmentation followed by speaker change detection and a speaker clustering and speaker verification. We analyze different clustering methods all based on a hierarchical classification technique. The experiments were conducted on a broadcast news data in Galician. The results show the potential of using a priori speaker models for some speakers involved in the conversation.

## 1. INTRODUCTION

Over the last years, a huge amount of measures and signals have been proposed and investigated for use in biometric recognition systems. Among the most popular features are fingerprint, face and voice. Recent advances and successes in speaker recognition technology have made it possible to use voice as a significant biometric feature. Thus, there are many application areas for speaker recognition technology. Such applications cover almost all the areas where it is desirable secure access or any type of interaction that need to identify or authenticate a person.

An emerging application area where speaker recognition technology is involved is the field of structuring the information of multimedia (audio-visual) documents. These multimedia documents contain, in general, multi-speaker audio recordings, and for some applications it may be relevant to determine *who spoke when*, also referred to as speaker segmentation and clustering in the literature. Thus, the segmentation of the data in terms of speakers could help in efficient navigation through audio documents like meetings recordings or broadcast news archives. Using these segmentation queues, an interested user can directly access a particular segment of the speech spoken by a particular speaker. Besides, in a forensic context, it could be necessary to track the time intervals where a suspect is talking in a lengthy telephone tap.

Traditionally, the speaker recognition task supposes that training and test are composed of mono-speaker records. Then, to handle this kind of multi-speaker recordings, some extensions of the speaker recognition task are needed, such as

- The N-speaker detection which is similar to speaker verification. It consists in determining whether a set of target speakers are speaking in a conversation.
- Speaker tracking that consists in determining if and when a target speaker speaks in a multi-speaker record.
- Speaker segmentation that is close to speaker tracking but there is no information about the identity and number of speakers present. The aim of this task is to determine the number of speakers and also when they speak. This problem corresponds to a blind classification of the data, and the result is a partition in which every class is composed of segments of one speaker.

In this paper, we focus on the problem of speaker segmentation, detection and tracking in multi-speaker audio recordings using speaker biometrics. With the work presented here, our aim was to investigate the performance of different clustering approaches involved in the speaker segmentation step, and also to investigate the interaction between the speaker segmentation and speaker detection modules in a task where the identity of some speakers is known *a priori*. In [1] the authors analyze the benefits of including different kinds of *a priori* information for speaker segmentation in a speaker diarization system.

The emphasis of this research work was on methods for speaker clustering. This involves research into statistical methods for data classification and clustering. Following the state-of-art speaker diarization systems, we decided to use hierarchical agglomerative classification approaches in the problem of speaker clustering, and to analyze how to measure the similarity between clusters if we only have the similarity between pairs of speech segments.

The paper is organized as follows. Section 2 gives a brief description of the speaker change detection system. Section 3 proposes some approaches for the speaker clustering task. In Section 4 we give an explanation of our experimental framework. The performance of the various clustering methods are shown and discussed in Section 5. Finally, Section 6 concludes this work and gives some future directions.

## 2. SYSTEM OVERVIEW

## 2.1 Acoustic segmentation and classification

To begin with, the audio stream should be broken down into homogeneous acoustic audio regions. So, the segmentation algorithm finds changes in the acoustic conditions and marks those time instants as segment boundaries. Then, each segment is classified as speech or non-speech, and the speech segments are classified according the speaker gender: male or female.

Most of the multimedia recordings have two related information sources, namely audio and video streams, which are often synchronized: acoustic changes occur more likely in the neighborhood of video shot boundaries. Having this into account, our acoustic segmentation module is based on a multimedia approach that uses not only audio but audio and video processing for the audio stream segmentation [2]. Such approach is based mainly on the Bayesian Information Criterion (BIC) [3].

This segmentation system works through three stages. In the first phase the BIC algorithm is applied using a window of 2 seconds. Here, the aim is to determine the acoustic change candidates. In the second phase, which is optional, these candidates are used as centers of short windows of 2 sec applied in a histogram-based shot boundary detector. So, information about the reliability of the presence of a shot boundary is given to the third phase. Finally, the last phase applies also the BIC algorithm but using a fixed size window (8 sec) centered in the candidates obtained in the first phase. In this phase, the information provided by the shot boundary detector is used to dynamically tune the BIC penalty weight, λ. If shot boundary is detected, the penalty weight is reduced otherwise its value remains unaffected. By reducing the penalty weight, the probability that the candidate has been accepted as a true acoustic change is increased. The factor by which the penalty weight is reduced depends on the reliability of the detected shot boundary.

### *2.1.1. Speech/Non-speech classification*

After the acoustic segmentation stage described above, each segment is classified using a speech/non-speech discriminator. This stage is very important for the rest of processing since we are not interested in processing audio segments that do not contain useful speech. We use an approach based on maximum likelihood classification with Gaussian Mixture Models (GMM). Specifically, five classes or models are considered: pure music, clean speech, noisy speech, noise and silence. Thus, the pure music, noise and silence segments are discarded.

### *2.1.2. Gender classification*

This module classifies each speech segment into one of two possible sex: male or female. This is done using also a GMM-based maximum likelihood classifier. Gender classification is used to improve the speaker clustering task. By separately clustering each gender class we will reduce the search space when evaluating the proposed hierarchical agglomerative clustering algorithm. It also avoids segments having opposite gender tags being erroneously clustered together.

## 2.2 Speaker change detection

Speaker change detection (SCD) solves the problem of finding speaker segment boundaries. It is also considered as a part of the problem that no information about the speaker identities and their number is known a priori. Several approaches have been proposed in the literature for SCD that can be broadly categorized as decoder based, model based and metric based approaches. The most common used approaches are metric based, mostly because these approaches have the advantage that they do not rely on any pre-existing models, which tend to fail in unseen conditions.

As described in section 2.1, in our system, the input audio stream is first processed by the acoustic segmentation module, which detects transitions in the audio characteristics. The module is expected to find change points whenever there is a remarkable change in the acoustic properties of the input signal. Thus, it will find change points whenever there is a transition from one speaker to another. However, the task in practice is a little more complicated as these transitions are not so obvious all the times. For example, in mostly broadcast news shows, the anchor starts speaking with music in the background and the intensity of this music gradually fades. In this situation the segmentation algorithm will find a change point when the background switches from background music to clean speech. Moreover, the environmental sounds and noise, sometimes, change while a reporter is speaking. This makes that the segmentation algorithm also finds a change point each time there is a change in the background acoustic condition.

In order to remove the change points corresponding to the above mentioned environment changes, and keep only the speaker changes, we will measure the similarity between adjacent speech segments. As similarity measure we propose the cross-likelihood ratio between the two speech segments, i.e., for segments i and j with models $\lambda_i$ and $\lambda_j$, this distance is defined as

$$d(i,j) = \log\left(\frac{p(i \mid \lambda_i)}{p(i \mid \lambda_j)}\right) + \log\left(\frac{p(j \mid \lambda_j)}{p(j \mid \lambda_i)}\right)$$

If d(i,j) falls below a threshold, the change point between both segments is removed.

Speaker change detection step is important for the next step, speaker clustering. If we falsely detect a speaker changing point, we can compensate for the error through the speaker clustering step. However, if we skip the real changing point, the clustering step cannot recover it. For these reasons, we tightly detect changing points to avoid the skip, although some points can be wrongfully detected as changing points.

## 2.3 Speaker clustering

The "clustering" process, in general, can be defined as unsupervised classification of data, i.e., without any a priori knowledge about the classes or the number of classes. In our task, the clustering process should result, ideally, in a single cluster for every speaker identity. We propose to use a *hierarchical agglomerative* clustering approach in order to group together segments from the same speaker.

First of all, we train an Universal 32 components GMM, Θ, based on all the speech segments in the audio stream. Then, the Θ model is MAP adapted in order to obtain a set of specific models $\theta_i$ representing each segment. Finally, a hierarchical classification algorithm is applied in three steps: 1) the first stage consists in computing some distance-like measure between each pair of segments; 2) the second step consists in creating a tree structure by starting with each segment in its own cluster, and recursively merging clusters according to some distance-related criterion; 3) and the last stage is to pick one of the partitions, a process we call tree cutting.

Several measures of similarity between segments or clusters can be used. Our first experiments were conducted using a distance measure which uses information about the likelihood score. Specifically, for each speech segment, $s_i$, it is scored against all the set of trained segment models, $\theta_i$, and the collection of scores is used to form the N-dimensional *segment characterization vector* (N segments). So, the similarity between two segments can be computed straightforwardly using the cosine distance between the two corresponding vectors. This distance ignores absolute sizes of the measurements, and only considers their relative ones. It is a popular distance measure for comparing documents in the information retrieval literature.

Hierarchical agglomerative methods are well documented in the literature [4]. The aim is to pick the closest pair of clusters according to the distance matrix, and merge them. This step is repeated until there is only one cluster. The distance matrix only gives the distance between pairs of single segments, so some method is required to construct a distance between clusters from distances between single segments. There are several possibilities most of them are variants of the single-link, complete-link, group average-link and minimum-variance algorithms. Among these, the single-link, complete-link and group average-link are the most popular. These algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete-link algorithm, the distance between two clusters is the maximum of all pairwise distances between patterns in the two clusters. In the group average-link approach, the distance between two clusters is the average of all pairwise distances between patterns in the two clusters. Very little is known about what qualities make a cluster distance good for clustering. Then, we propose to analyze the performance of these three approaches in our task of speaker clustering.

Several techniques exist in the literature for selecting a partition. These techniques consist in cutting the tree at a given level (or height) or in pruning the tree by selecting clusters at different levels. In our experiments, we selected the partition by cutting the tree based on a specified maximum number of clusters. This number of clusters will be chosen based on the knowledge of the stimation of the number of soeaker number of speakers in the audio file.

## 2.4 Speaker detection

It is the last stage in which we attempt to identify those speaker clusters that were spoken by a set of the pre-defined news anchors. We can consider two different methods. One of which labels each speech segment as one of the known speakers or as an unknown speaker, before the clustering stage. And the other one labels each speaker cluster as one of the known speakers or as an unknown speaker.

To detect or identify the known speakers we will use the common techniques used in speaker verification tasks. We compute the likelihood ratio between the known speaker and a background model, and compare this score with a threshold.

### 3. EXPERIMENTAL FRAMEWORK

News broadcasting is an application where segmenting the audio stream by speaker identities may be interesting for many purposes. In the last years we have been working in Automatic Transcription of Broadcast News (BN). For that research work we have collected a corpus of TV broadcast news in Galician. Then, the algorithms described in this paper were evaluated on this BN database. The corpus consists of 31 news shows. Each show is about 60 minute long, and there are about 75 speakers and about 170 speaker changes (turns) in each file. The corpus is divided into a training set (26 news shows), a development set (2 news shows) and a test set (3news shows). The speaker segmentation, detection and tracking are performed on the test set.

Often, in news broadcasting we have knowledge about the anchors and therefore we can have trained anchor models a priori. In our BN database the number of known target speakers (anchors) is three (1 male speaker and 2 female speakers). Due to the fact that the number and some of the speaker identities are not known it is desirable to perform unsupervised speaker segmentation.

## 3.1 System configuration

The audio signal is characterized by 12 mel-frequency cepstral coefficients extracted every 10 ms using 25 ms hamming windows. Then these cepstral features are augmented by the $0^{th}$ cepstrum coefficient and the log-energy. The acoustic change detection stage uses only the 12 MFCCs and the $0^{th}$ cepstrum coefficient as features. In the speech/non-speech classification and the gender classification modules the first derivatives of this 13-dimensional feature vector were also considered. In the clustering stage CMS was applied in order to remove channel effects.

The speech, non-speech, male and female models were 32 diagonal GMM directly trained on data extracted from the train corpus using the EM algorithm.

The anchor models were also 32 diagonal GMM directly trained on data extracted from the train corpus using the EM algorithm. The data available for every anchor was between 20 and 60 minutes.

## 4. EXPERIMENTAL RESULTS

### 4.1 Evaluation measures

There are two kinds of errors to be considered: the speech detection errors and the speaker mapping errors:

- Miss speech: a segment was incorrectly labeled as non-speech.
- False alarm speech: a segment was incorrectly labeled as speech.
- Speaker error: a segment was assigned to the wrong speaker.

We will describe the performance of the speaker change detection module using the measures of Recall (% of detected speaker change points), Precision (% of detected points which are genuine change points) and F-measure (defined as 2*Recall*Precision/(Recall+Precision)).

In order to evaluate the clustering, a bi-directional one-to-one mapping of reference speakers to clusters is computed. The mapped speaker-cluster pairs define the correct cluster for the speaker and the correct speaker for the cluster. Using the correct speakers/clusters, the *Q-measure* is defined as the geometrical mean of (1) the percentage of speaker frames referring to the correct speaker and (2) the percentage of speaker frames labeled with the correct cluster. Another performance measure (related to Q) is the Diarization Error Rate (DER). It is defined by NIST as the percentage of frames with an incorrect cluster-speaker correspondence.

### 4.2 Results

Results of the speech/non-speech module are summarized in table 1. The performance measures are the percentages of speech frames and non-speech frames that are classified correctly. The F-measure between them represents the accuracy of the speech/non-speech classification task. We can see that the errors are mainly in missing non-speech segments. This will increase the error on the next modules.

| Speech | Non-Speech | Accuracy |
|---|---|---|
| 99.1 | 84.2 | 98.5 |

**Table 1**: Results for speech/non-speech classification.

Table 2 presents the performance of the speaker change detection module. These results are good according to the state-of-art systems. We can see as the recall is higher than the precision, which reflected a high number of false alarms. Some of the errors were due to the errors of the speech/non-speech detector.

| Recall | Precision | F-measure |
|---|---|---|
| 90.0 % | 71.7 % | 79.82 % |

**Table 2**: Results for speaker change detection.

In table 3 we show the Q-measure and the DER measure obtained after the clustering algorithm using different agglomerative methods. The DER figures include the FA/Miss speech errors. We can see that the method that gives better performance is the complete link method. This method is the one that gives more compact clusters.

| Agglomerative method | Q-measure | DER |
|---|---|---|
| Single-link | 63.8 | 26.2 |
| Complete-link | 80.1 | 17.5 |
| Average-link | 73.8 | 20.5 |

**Table 3**: Results for the clustering process.

Table 4 shows the results obtained in the anchor detection task. Results are given in terms of percentage of deletions, that is, clusters not identified as belonging to the anchor, and percentage of insertions, that is clusters incorrectly labeled as anchor. The results are very promising especially due to the very low insertion rate. These results were obtained using the method of labeling after the clustering stage. The other case, labeling after speaker change detection, will be investigated in the future work.

| Anchor | deletions | Insertions |
|---|---|---|
| Female 1 (news) | 1.3 % | 2.2 % |
| Female 2 (sports) | 4.6 % | 1.9 % |
| Male 1 (weather) | 1.7 % | 1.4 % |

**Table 4**: Results for the anchor identification process.

## 5. CONCLUSIONS AND FUTURE WORK

This paper reports our work on the development of an audio segmentation system which uses a priori information about a set of speaker identities for speaker detection and tracking. The results show that a speaker-indexing system can help in a speaker recognition task where there is more than one speaker in the conversation. These results are good according to the state-of-art systems [5].

We have investigated different agglomerative methods in the hierarchical speaker clustering stage. Very little is known about which method is the best and it is depending on the database and the task. We will continue our research in this field using other distance measures between segments, and also using other criteria to group segments based on the analyzed in this work.

## 6. ACKNOWLEDGEMENTS

This paper has been partially supported by Spanish MEC under the project TIC2002-02208, and Xunta de Galicia under the projects PGIDT03PXIC32201PN, PGIDT02TIC32201PR.

## 7. REFERENCES

1. D. Moraru, L. Besacier, E. Castelli. “Using a priori information for speaker diarization,” in Proc. Odyssey 2004, Toledo (Spain).
2. L. Perez-Freire and C. Garcia-Mateo. “A multimedia approach for audio segmentation in TV broadcast news,” in Proc. ICASSP 2004.
3. S. Chen, P.S. Gopalakrishnan. “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” in DARPA Proc. Speech Recognition Workshop, 1998.
4. A.K. Jain, M.N. Murty and P.J. Flynn, “Data clustering: A review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999.
5. NIST, “Benchmark tests: Rich Transcription (RT),” http://www.nist.gov/speech/tests/rt/.

## COST

**COST is an intergovernmental European framework for international cooperation between nationally funded research activities. COST creates scientific networks and enables scientists to collaborate in a wide spectrum of activities in research and technology. COST activities are administered by the COST Office.**

**COST is supported by the EU Framework Programme**