MMUA

Proceedings

Workshop on
Multimodal User Authentication

December 11-12, 2003
Santa Barbara, CA

Supported by:

UC Discovery Grant, Industry-University Cooperative Research Program

France Télécom

In cooperation with:

EURASIP

IEEE Signal Processing Society

ISCA

2003 Workshop on Multimodal User Authentication (MMUA 2003)

Table of Contents

2003 Workshop on Multimodal User Authentication (MMUA 2003)

Organizing Committee

Jean-Luc Dugelay, Institut Eurécom, France
Jean-Claude Junqua, Panasonic Speech Technology Laboratory (PSTL), USA
Kenneth Rose, University of California, Santa Barbara (UCSB), USA
Matthew Turk, University of California, Santa Barbara (UCSB), USA

Program Committee

Bir Bhanu, UC Riverside, USA
Josef Bigun, Halmstad University, Sweden
Jean-François Bonastre, University of Avignon, France
Hervé Bourlard, IDIAP, Switzerland
Rama Chellappa, University of Maryland, USA
Gérard Chollet, ENST, France
Farzin Deravi, University of Kent, U.K.
Jana Dittmann, IPSI, Germany
Bernadette Dorizzi, INT-GET, France
Touradj Ebrahimi, EPFL, Switzerland
Sadaoki Furui, Tokyo Institute of Technology, Japan
Larry Heck, Nuance, USA
Javier Hernando, UPC, Spain
Anil Jain, Michigan State University, USA
Kenneth K. M. Lam, Hong Kong Poly. Univ. Hong Kong
B.S. Manjunath, UCSB, USA
John Mason, Swansee University, UK
Philippe Morin, PSTL, Panasonic, USA
Javier Movellan, UCSD, USA
Kenji Nagao, ATRL, Panasonic, Japan
Shrikanth Narayanan, USC, USA
Javier Ortega-Garcia, Tech. Univ. of Madrid, Spain
Sharon Oviatt, OHSU, USA
Ioannis Pitas, Thessaloniki University, Greece
Salil Prabhakar, DigitalPersona, USA
Douglas Reynolds, MIT Lincoln Lab., USA
Stefano Soatto, UCLA, USA
Doug Tygar, UC Berkeley, USA
James Wayman, SJSU, USA

Introduction and Welcome

The proliferation of information access terminals, coupled with the increasing use of information sensitive applications, such as electronic commerce and health care, has triggered a real need for reliable, user-friendly, and commonly acceptable control mechanisms for accessing private and confidential information. The goal is to protect the individuals who use such applications as well as the organizations offering them. The conventional means of identity verification for access control such as passwords, personal identification numbers, passports, and identification cards can easily be compromised. In view of this, it appears that the required level of reliability in determining the identities of individuals may only be achieved through the use of biometrics.

Many applications concentrate on one biometric modality only (for example, fingerprint or iris-scan) due to their high discrimination power. However, the suitability of each modality to a given application depends on various factors including the attitudes of users and their personalities as well as the operational environments and conditions. Authentication systems that are required to be robust in natural environments (e.g., in the presence of noise and illumination changes) cannot rely on a single modality. In addition, a single modality is not always appropriate, convenient, or available. Thus, fusion with other modalities is essential. Successful integration of multiple biometric modalities must be based on a thorough understanding of the individual sensing technologies and modalities and of their interaction.

Developing such systems requires advances in many different recognition and verification technologies, including those based on analyzing speech, vision, and behavior. Most importantly, advancement in this field requires the creation of a community of researchers willing to work in an interdisciplinary manner going beyond the well-established research communities. Speech researchers, for example, need to go beyond their traditional area of expertise and interact with computer vision or human interface researchers. Multimodal databases have to be collected and interdisciplinary research needs to be pursued. Finally, careful evaluation and assessment of multimodal systems has to take place.

The interdisciplinary nature of multimodal user authentication led us to organize this workshop with the specific goal of providing a forum for researchers from different disciplines to help establish collaborations and partnerships and to promote the sharing of information and cross-discipline research. The workshop is supported by a University of California Discovery Grant from the Industry-University Cooperative Research Program and a sponsorship from France Télécom R&D. It is held in cooperation with the International Speech Communication Association (ISCA), EURASIP and the IEEE Signal Processing Society.

We would like to give special thanks to the three invited speakers (Josef Bigun, Gary Strong and James Wayman), who all accepted with enthusiasm the challenge of preparing overview talks, and to Jonathon Phillips (and the panelists) for organizing the panel discussion. Finally, our gratitude goes to the workshop scientific contributors and the members of the International Scientific Committee for their help in reviewing the submitted papers. It is also our pleasure to acknowledge and thank Tim Robinson from the University of California, Santa Barbara, for handling the workshop submissions and for his precious help in communicating with the authors and other important tasks.

We hope that this workshop will be successful from both a technical and social point of view and that the contacts and discussions you will have will be beneficial for your future research or business. Meanwhile, be sure to enjoy the beauty of Santa Barbara.

The Workshop Organizing Committee:
Jean-Luc Dugelay, Jean-Claude Junqua, Ken Rose, Matthew Turk

Invited Speaker

Gary Strong
U.S. Department of Homeland Security

Signature with Text-Dependent and Text-Independent Speech for Robust Identity Verification

B. Ly-Van*, R. Blouet**, S. Renouard**, S. Garcia-Salicetti*, B. Dorizzi*, G. Chollet**
** INT, dépt EPH, 9 rue Charles Fourier, 91011 EVRY France;*
***ENST, Lab. CNRS-LTCI, 46 rue Barrault, 75634 Paris*
Emails: {Bao.Ly_van, Sonia.Salicetti, Bernadette.dorizzi}@int-evry.fr;
{Blouet, Renouard, Chollet}@tsi.enst.fr

Abstract

This article addresses the setting up of a Biometric Authentication System (BAS) based on the fusion of two user-friendly biometric modalities: signature and speech. All biometric data used in this work were extracted from the BIOMET multimodal database [1]. The Signature Verification system relies on Hidden Markov Models (HMMs) [2], and we use two kinds of Speaker Verification systems. The first one is text-dependent and uses Dynamic Time Warping (DTW) [3] to compute a decision score. The second one is text-independent and based on Gaussian Mixture Models (GMMs) [4]. We first present the BIOMET database and describe precisely the two modalities of interest before giving a presentation of each monomodal BAS as well as their performance evaluation. We then compare performances of two classical learning-based fusion techniques: an additive CART-trees [5] classifier built with boosting [6], and Support Vector Machines (SVMs) [7]. In particular, the signature modality was fused with clean and noisy speech, at two different levels of degradation. The impact of noise in fusion performance is studied relative to that of each of the speech experts alone.

1 Introduction

Many commercial applications require a step of Identity Verification before accessing to a service or to sensitive data. As the media and channels through which the Identity Verification process takes place are becoming more diverse, multimodal biometric authentication systems could be used with convenience to improve user security. Moreover, several studies have already proven that combining different biometric modalities significantly improves the performances compared to system working with a single modality [8, 9]. We present in this article a bi-modal biometric system based on two well-accepted modalities: signature and speech. The two main virtues of those modalities are their physical non-intrusiveness and their capabilities to be easily sampled by personal computers or common electronic devices. Indeed, smart phones, tablet PC and Personal Digital Assistant (PDA) already allow the use of these two biometric modalities.

Speaker Verification systems usually work either in text-dependent or text-independent mode. In this paper, we use these two Speaker Verification (SV) working modes along with the on-line hand-written signature to set up our multimodal BAS. Indeed, these two SV modes may be very complementary for many applications. For example, during a phone access to sensitive data the text-dependent system can focus on keywords while the text-independent system works on the whole client utterance. Therefore, we finally have three different biometric systems.

We perform score fusion of those 3 systems by means of two different learning-based techniques: an additive CART-tree [5] classifier built with boosting [6], and Support Vector Machines (SVMs) [7]. SVMs have already been successfully used on multimodal biometric data [9, 10]. They have proven to be a powerful tool for classification, and well-suited for applications in which few data is available, as it is the case in identity verification. Also, decision trees have successfully been used to fuse the scores of biometric experts, for example in [11]. Moreover, boosting is known to be a very efficient tool to fit additive tree based classifiers [12]. It significantly improves performance and is also well-suited for applications with few training data. Hence, we decided to use boosting to fit an additive CART-tree classifier for multimodal fusion purposes. Finally, we propose a comparison of these two fusion paradigms on the BIOMET data, as well as a test of their robustness in the presence of noise in part of this data. Indeed, the capture of speech in a real life application is often done in noisy conditions. It is therefore important to study the impact of noise in expert scores fusion as recently done in [13].

This paper is organised as follows: section 2 describes the speech and signature data from the BIOMET database. Section 3 gives the principles of the signature verification expert, detailed in [14], as well as related experimental results. Section 4 describes the two speech verification experts. Fusion by additive CART-tree classifier and by SVM are studied in Section 5, with clean speech data first, and then with degraded speech data.

2. BIOMET brief description

BIOMET is a multimodal biometric database including face, image, finger print, signature and voice. We exploit signature and voice data from 68 people with time variability, captured in the two last BIOMET acquisition campaigns, which have a five months spacing between them. More details on the BIOMET database can be found in [1].

2.1 Signature data

The digitizer captures from each signature, at a rate of 200 samples per second, 5 parameters, including the coordinates of each point sampled on the trajectory *(x(t), y(t)),* the axial pen pressure *p(t)* in such a point, and the position of the pen in space (azimuth and altitude angles). The total number of genuine signatures available per person is 15 and 12 impostor signatures, made by four different impostors.

2.2 Speech data

Both speech sessions of the BIOMET database were recorded in quiet environment and using the same kind of microphone. Sampling rate is 16 kHz and sample size is 16 bits. In both sessions, each speaker uttered twice the 10 digits in ascending and descending order before reading sentences. The amount of available speech for each speaker is about 90 seconds by session.

3. Signature verification

3.1 Pre-processing and encoding signatures

There is noise in the data, on one hand due to the parameter quantification performed by the digitizer, and on the other hand its high sampling rate. Different filtering strategies were thus chosen according to each parameter, as motivated in [14]. Finally, 12 dynamic parameters are extracted on each point of the signature.

3.2 Modelling signatures

As we have few signatures for training a signer's HMM, we used Bagging [15] to produce an "aggregated HMM" of each signer's characteristics. Indeed, by combining models learned on different samplings of a given data set, one builds a model with a more complex and more stable output. We generated *T* different training sets, by sampling with replacement from the *N* original signatures at disposal for training purposes. We thus built *T* component models, which are *T* continuous left-to-right HMMs [2] with 2 states and 3 gaussians per state. We then computed a composite score *S(O)* on signature *O*, by averaging the *T* output scores obtained from the component models when a signature is presented.

We then built a classifier as follows: in order to decide whether the claimed identity of signer *i* is authentic or not, we compute the absolute difference of the composite score $S_i(O)$ of his/her current signature *O* and the average value S_i* of the *T* component models' output scores on their respective training data set (we have *T* data sets); finally, we compare this quantity to a threshold. Indeed, a signature *O* is accepted if and only if:

$$| S_i(O) - S_i^* | < \tau \qquad (1)$$

where τ is a global threshold, computed once for all signers, on a devoted database, as explained in Section 3.3.

3.3 Experimental setup

In order to train each signer's aggregate model, all the signatures of the third campaign are used (*N=10* signatures most of the time per person). Following the fact that the improvement of bagging is evident within ten replications of the original training set [15, 16], we chose *T=10* as the number of component models to be used.

Among each signer’s genuine signatures, 10 out of 15 are used to train the corresponding aggregate model, as described above. The remaining 5 genuine signatures and the 12 impostor signatures may be devoted to compute the global threshold τ in (1), or to test the system, according to the signer's number in the database: indeed, the database of 68 clients was split in two databases *BA* and *BT* of 34 clients each: *BA* to compute the threshold τ, and *BT* to test the system once the threshold has been computed.

The optimal threshold is computed on *BA* following two criteria: the Equal Error Rate *(EER)* corresponding to *FA = FR, FA* being the False Acceptance Rate and *FR* the False Rejection Rate, and the minimum of the Total Error Rate *(TE),* that is the number of errors made by the system (of type *FA* and of type *FR*), over the total number of signatures (genuine signatures and forgeries as well) presented in *BA*. As we use an aggregate model for each signer, the optimal threshold is found on database *BA* as follows: for each possible value of τ, for each signature *O* belonging to signer *i* of *BA*, the corresponding composite score *S(O)* is computed by averaging the *T=10* scores of the *T* component models of signer *i*. Then the decision of acceptance or rejection is taken according to (1).

The system is then tested on *BT*. Table 1 shows the performance obtained for both criteria *EER* and *TE* on such database, with the corresponding 95% confidence interval [17].

Table 1. Global performances of signers' aggregate models

Criterion	TE (%)	FA (%)	FR (%)
EER	11.1 [±2.6]	9.5 [±3.0]	14.8 [±5.4]
Minimum TE	11.9 [±2.7]	8.9 [±2.9]	20.1 [±6.0]

Roughly, we notice that the signature expert presents a Total Error Rate of around 10% (with both criteria *EER* and Minimum *TE*), with a rather large confidence interval. This result can be explained by the low number of samples available in the BIOMET database compared to other on-line signature databases [18]. Generally, this difficulty is indeed inherent to personal identity verification applications: one can hardly imagine building very large databases of biometric data for each application. Also, the signature modality, contrary to other biometric modalities, has the particularity of forgeries that are made by impostors that intentionally imitate the genuine signatures which increases this difficulty.

4. Speech verification

4.1 Introduction

Speaker Verification systems decision is mostly based on a simple hypothesis test between two hypotheses H_λ and $H_{\bar{\lambda}}$ with:

H_λ : X has been uttered by λ

$H_{\bar{\lambda}}$: X has been uttered by another speaker

Hence, the score is usually based on two similarity measures and the claimed identity is confirmed according to:

$$\log \frac{\mathbf{D}_\lambda(\mathbf{X})}{\mathbf{D}_{\bar{\lambda}}(\mathbf{X})} \left\} \begin{array}{l} \geq \boldsymbol{\beta} \textbf{ accept } \boldsymbol{\lambda} \\ < \boldsymbol{\beta} \textbf{ reject } \boldsymbol{\lambda} \end{array} \right. \quad (2)$$

where $D_\lambda(X)$ and $D_{\bar{\lambda}}(X)$ are respectively the similarity measures of the speech utterance *X* conditionally to H_λ and $H_{\bar{\lambda}}$ and *β* is the decision threshold. As described in Section 4.2, the text-dependent Speaker Verification system relies on Dynamic Time Warping (*DTW*) [3] to compute $D_\lambda(X)$ and $D_{\bar{\lambda}}(X)$. In the text-independent approach, as described in Section 4.3, $D_\lambda(X)$ and $D_{\bar{\lambda}}(X)$ respectively correspond to the probability density functions $P_\lambda(X)$ and $P_{\bar{\lambda}}(X)$ associated to the densities of H_λ and $H_{\bar{\lambda}}$ given *X*. The state-of-the-art approach consists in using Gaussian Mixture Models (GMMs) [4] to estimate those densities.

The same kind of acoustic analysis is used in the text-dependent and text-independent approach. Every 10ms, we first extract from each 20ms frame of speech a 32 dimensional acoustic vector composed of 16 mel-scale filter bank cepstral coefficients augmented by associated delta coefficients. Delta cepstra are computed over ± 2 feature vectors.

4.2 Text-dependent Speaker Verification

In the text-dependent Speaker Verification system, the decision score is based on the *DTW* [3] distance between the training sequence X_λ of 4 digits with an utterance X of the same sequence of digits. As in [19], we use a cohort of speakers to compute $D_{\bar{\lambda}}(X)$. For each client λ, the cohort is composed of a set $\Gamma_\lambda = \{X_1 .... X_K\}$ of *K* speech segments of speakers uttering the same sequence of digits. $D_{\bar{\lambda}}(X)$ is the mean over $\Gamma_\lambda$of the log-*DTW* distance between *X* and X_k with $k=1...K$. $D_\lambda(X)$ corresponds to the log-*DTW* distance between *X* and X_λ . The decision score for a test sequence corresponds to the subtraction of $\mathbf{log}(\mathbf{D}_\lambda(\mathbf{X}))$ with $\mathbf{log}(\mathbf{D}_{\bar{\lambda}}(\mathbf{X}))$.

4.3 Text-independent Speaker Verification

In the text-independent Speaker Verification system, we use a single speaker-independent model to represent $P_{\bar{\lambda}}(X)$. This model, also called UBM [4], corresponds to a 256 components GMM with diagonal covariance matrices. Each client model is obtained by a mean-only Bayesian adaptation of the UBM [4] using associated training speech data. The decision score for a test sequence corresponds to the mean log-likelihood ratio computed on the whole test utterance.

4.4 Experiments on speech data

4.4.1 Evaluation protocol. In both text-dependent and independent Speaker Verification systems, the client or target speaker set is composed of 68 speakers from the BIOMET database. For the text dependent system, the training data for a target speaker is one utterance of 4 digits (about 2s of speech). The cohort of speakers is composed of 50 utterances of the same digits. Test data is composed of 5 genuine accesses and 12 impostor accesses. In the text-independent system, $P_{\bar{\lambda}}(X)$ is trained using the whole speech data available in the BIOMET database (about 4 hours of speech). Half of these 4 hours of speech are uttered by speakers that are not impostors nor clients. Each client model is adapted from the speaker using the 10 digits utterance (about 15s of speech). Test data is composed of a segment of speech of approximately 15s, taken from read utterances. The training speech material is based on digit vocabulary and the test speech material is based on uttered word. For each speaker we performed 5 genuine and 12 impostor accesses. Both systems have been evaluated under 3 different conditions of noise in test utterances: without noise, with a gaussian white noise of -10dB, and with a gaussian white noise of 0dB.

4.4.2 Results. Performances of text-independent and text-dependent Speaker Verification systems are given respectively in Table 2 and Table 3.

Table 2. Performances of the text-independent Speaker Verification system

SNR	Criterion	Error (%)	FA (%)	FR (%)
without noise	EER	7.3 [±2.2]	5.8 [±2.4]	10.7 [±4.7]
	Min. TE	6.3 [±2.0]	2.0 [±1.4]	16.0 [±5.5]
10 dB	EER	12.0 [±2.7]	13.2 [±3.4]	9.5 [±4.4]
	Min. TE	8.0 [±2.3]	2.0 [±1.4]	23.2 [±6.4]
0 dB	EER	29.4 [±3.8]	34.0 [±4.8]	19.0 [±5.9]
	Min. TE	17.0 [±3.1]	6.0 [±2.4]	45.0 [±7.5]

Table 3. Performances of the text-dependent Speaker Verification system

SNR	Criterion	Error (%)	FA (%)	FR (%)
without noise	EER	13.5 [±2.9]	16.4 [±3.7]	7.1 [±3.9]
	Min. TE	10.3 [±2.6]	7.6 [±2.7]	17.0 [±5.7]
10 dB	EER	16.0 [±3.1]	19.8 [±4.0]	7.7 [±4.0]
	Min. TE	11.9 [±2.7]	7.8 [±2.7]	22.1 [±6.3]
0 dB	EER	21.2 [±3.4]	25.3 [±4.4]	11.8 [±4.9]
	Min. TE	16.5 [±3.1]	6.3 [±2.4]	42.0 [±7.4]

5. Fusion

5.1 Additive Tree Classifier

Boosting permits to construct efficient additive modelization from a so-called weak learner. This weak-learner here corresponds to a classical binary tree built with the CART [5] algorithm. This algorithm permits one to construct a tree by recursive split of the observation space, here corresponding to the 3-D scores space of the signature modality expert and both speech verification experts.

For instance, as shown in Figure 1, R^k is split in $R^{k,l}$ and $R^{k,r}$ when maximizing ΔH:

$$\Delta H = H(R^k) - p_l H(R^{k,l}) - p_r H(R^{k,r})$$

where $H(R^k)$, $H(R^{k,l})$ and $H(R^{k,r})$ are entropies of

nodes R^k, $R^{k,l}$ and $R^{k,r}$ with:

$$H(R) = p_\lambda(R) \cdot \log\left(p_\lambda(R)\right) + p_{\bar{\lambda}}(R) \cdot \log\left(p_{\bar{\lambda}}(R)\right)$$

$$p_l = \frac{N^{k,l}}{N^k},\ p_r = \frac{N^{k,r}}{N^k} \text{ and } p_\lambda(R) = \frac{N_\lambda(R^k)}{N(R^k)}$$

in which $N^{k,l}$, $N^{k,r}$ and N^k are respectively the number of observations in nodes $R^{k,l}$, $R^{k,r}$ and R^k, and $N_\lambda(R^k)$ is the number of observations of class λ in R^k.

In our experiments, a node R^k is split only if $N^k > 50$. The score S_i associated to each vector $s=[s_1, s_2, s_3]$ is

$$S_i = \log \frac{p(\lambda|s)}{p(\bar{\lambda}|s)}, \quad \text{with} \quad p(\lambda|s) = p_\lambda(R) \quad \text{and}$$

$p(\bar{\lambda}|s) = p_{\bar{\lambda}}(R)$ if s is affected to the region R by the tree.

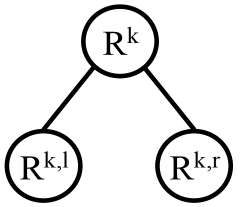


Figure 1. An additive tree classifier. The observation space R^k is split into two subspaces $R^{k,l}$ and $R^{k,r}$.

Given CART, a one-tree building algorithm, we use RealAdaboost [20] to fit an additive model. In this iterative algorithm, observations that have been incorrectly classified by the previous trees in the training ensemble are resampled with higher probability, leading to a new probability distribution for the next training ensemble.

The fusion decision score S is then obtained as the mean over all trees of S_i.

5.2 Support Vector Machines

In few words, SVMs' goal is to look for a hyperplane in a large dimension space which is considered because the input data are not linearly separable in the original space. We maximize the distance between the surface and the data, which leads to good generalization performance. Let $X=(x_i)$ be the data with labels $Y=(y_i)$ where $y_i = +1$ or -1 represents the class of each person, and Φ is the function which sends the input data X in the feature space F. The distance between the hyperplane

$H(w,b) = \{x \in F : \langle w , x \rangle + b = 0\}$

and X is called the margin Δ. Following the Structural Risk Minimization (SRM) principle, Vapnik [7] has shown that maximizing the margin (or minimizing $||w||$) leads to an efficient generalization criterion. One defines in F the kernel K as:

$K(x,y) = \langle\Phi(x), \Phi(y)\rangle$

Thanks to this function, we avoid handling directly elements in F. The optimal hyperplane is found by solving, as shown in [7], a quadratic convex problem and, from the optimality conditions of Karush-Kuhn-Tucker, one can rewrite w in the following condensed manner:

$$w = \Sigma_{i \in SV}\, \alpha_i\, y_i\, \Phi(x_i) \qquad (3)$$

where $SV = \{i\colon \alpha_i > 0\}$ denotes the set of support vectors.

The choice of Φ or equivalently K is very important in order to obtain an efficient solution. Traditionally, one chooses the Vapnik polynomial kernel $K(x,y)=\langle\Phi(x), \Phi(y)\rangle^d$ or the Gaussian kernel $K(x,y)=exp(-\gamma||x-y||^2)$. We have chosen a linear kernel ($d = 1$). Indeed, the use of this type of kernel in a similar fusion case [8] gave better performance, compared to other choices.

We will fuse the scores of the three experts, each designed for the same person. We thus put at the SVM three inputs, one per expert. The first one, for the signature modality, given signature O, is:

$$(S_i(O) - S_i{*})/\sigma \qquad (4)$$

where $S_i(O)$ and $S_i{*}$ are defined in Section 3.2; σ is the average of the standard deviations $\sigma(i)$ computed for person i in *FLB* as follows: we consider the scores given by the T component models of person i on the T corresponding genuine signatures data sets generated for bagging; and we compute their standard deviation $\sigma(i)$.

The second and third inputs to the SVM are the quantities log $\{D_{\lambda}(X)/D_{\bar{\lambda}}(X)\}$ in equation (2), where λ and $\bar{\lambda}$ are respectively estimated in text-independent and text-dependent modes described in sections 4.2 and 4.3.

5.3 Experiments

5.3.1. Fusion database. Following the same protocol as the one of the signature framework, we split the database of 68 persons in 2 subsets of 34 persons each, respectively named *FLB* (Fusion Learning Base) and *FTB* (Fusion Test Base). For each person in *FLB* and *FTB*, we have in general at disposal 5 genuine bimodal values and 12 imitation bimodal values.

5.3.2. Results. Table 4 presents the results of the different verification systems (Signature, Text-independent (TI) Speech, and Text-dependent (TD) Speech) as well as the results of the two fusion systems (Additive Tree Classifier (ATC) and SVM) for individuals of *FTB*, with the corresponding 95% confidence interval. These results have been obtained through a minimization of the global error rate *TE*.

Table 4. The performance of the fusion systems

	Model	TE (%)	FA (%)	FR (%)
	Signature	11.9 [±2.7]	8.9 [±2.9]	20.1 [±6.0]
Speech without noise	TI Speech	6.3 [±2.0]	2.0 [±1.4]	16.0 [±5.5]
	TD Speech	10.3 [±2.6]	7.6 [±2.7]	17.0 [±5.7
	ATC	2.8 [±1.4]	1.7 [±1.3]	5.2 [±3.3]
	SVM	2.7 [±1.4]	1.3 [±1.1]	5.9 [±3.6]
SNR: 10dB	TI Speech	8.0 [±2.3]	2.0 [±1.4]	23.2 [±6.4]
	TD Speech	11.9 [±2.7]	7.8 [±2.7]	22.1 [±6.3]
	ATC	2.9 [±1.4]	2.5 [±1.6]	3.9 [±2.9]
	SVM	2.9 [±1.4]	1.9 [±1.4]	5.3 [±3.4]
SNR: 0dB	TI Speech	17.0 [±3.1]	6.0 [±2.4]	45.0 [±7.5]
	TD Speech	16.5 [±3.1]	6.3 [±2.4]	42.0 [±7.4]
	ATC	6.7 [±2.1]	4.7 [±2.1]	11.2 [±4.8]
	SVM	5.8 [±2.0]	2.4 [±1.5]	13.6 [±5.2]

Roughly, we notice that in all cases, fusion reduces error rates of the best monomodal system by a factor 2. Also, it appears that the ATC and the SVM are equivalent in these experiments, in all the configurations here considered (clean or noisy environments). Finally, both fusion systems here studied show a good robustness to noise.

6. Conclusions

In this article, we have shown that the use of data fusion allows to improve significantly the performance of three unimodal identity verification systems. Indeed, we had at our disposal one signature and two speaker verification systems. We compare an Additive Tree Classifier (ATC) and a SVM on the BIOMET multimodal database and also study their robustness to the presence of noise in speech data. Two levels of degraded speech data were considered. It appears that the ATC gives very good results, equivalent to those of the SVM, and that in clean or noisy environments. This shows the importance of the boosting algorithm here used to build the ATC. Also, both fusion systems are resistant to the presence of noise. Indeed, in the best conditions, the Total Error Rate is around 2.8% for both fusion systems, and this rate is hardly lowered (to 2.9%) in the presence of noise at -10 dB. These results are encouraging, since few data is used to train the fusion systems.

7. References

[1] S. Garcia-Salicetti, C. Beumier, G. Chollet, B. Dorizzi, J. Leroux-Les Jardins, J. Lunter, Y. Ni, D. Petrovska-Delacretaz, "BIOMET: a Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities", *4th International Conference on Audio and Vidio-Based Biometric Person Authentication*, 2003.

[2] L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", *Prentice Hall Signal Processing Series*, 1993.

[3] Furui S., “Cepstral Analysis Technique for Automatic Speaker Verification”, *IEEE Trans. Acoustic, Speech, Signal Processing*, Vol ASSP – 29, 254-272, 1981.

[4] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, Vol. 10, No. 1, pp. 19-41, Jan. 2000.

[5] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, “Classification and Regression Trees”, *Belmont*, CA: Wadsworth*,* 1984.

[6] Y. Freund, “Boosting a Weak Learning Algorithm by Majority”, *Proceedings of the Third Workshop on Computational Learning Theory*, Morgan-Kaufman, 202-216, 1990.

[7] V. Vapnik, “The Nature of Statistical Learning Theory”, *Statistics for Engineering and Information Science,* Second Edition, Springer, 1999.

[8] S. Ben-Yacoub, "Multi-Modal Data Fusion for Person Authentification using SVM", *IDIAP Research Report 98-07*, 1998.

[9] S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, “Fusion of Face and Speech Data for Person Identity Verification”, *IEEE Trans. On Neural Networks*, Vol. 10, No 5, 1999, pp. 1065-1074.

[10] B. Gutschoven, P. Verlinde, “Multimodal Identity Verification using Support Vector Machine”, *Fusion 2000*, 2000.

[11] Arun Ross, Anil Jain and Jian-Zhong Qian, “Information Fusion in Biometrics”, *3rd Int'l Conference on Audio- and Video-Based Person Authentication,* AVBPA*,* pp. 354-359, Sweden, June 6-8, 2001.

[12] H. Drucker and C. Cortes. “Boosting decision trees”. *Advances in Neural Information Processing Systems*, volume 8. 1996.

[13] C. Sanderson, K. K. Paliwal, “Information Fusion and Person Verification using Speech and Face Information”, *IDIAP Research Report, 02-33*, September 2002.

[14] M. Fuentes, S. Garcia-Salicetti, B. Dorizzi "On-line Signature Verification: Fusion of a Hidden Markov Model and a Neural Network via a Support Vector Machine", *IWFHR8*, August 2002.

[15] L. Breiman, "Bagging predictors", *Machine Learning*, 24(2), pp. 123-140, 1996.

[16] J.R. Quinlan, "Bagging, Boosting, and C4.5", *Proceedings of the 13th National Conference on Artificial Intelligence*, pp. 725-730, 1996.

[17] P. Verlinde, “A Contribution to Multimodal Identity Verification Using Decision Fusion”, *Ph.D. Thesis*, Department of Signal and Image Processing, Telecom Paris, France, 1999.

[18] J.G.A. Dolfing, “Handwriting Recognition and Verification, a Hidden Markov Approach”, *Ph.D. Thesis*, Philips Electronics N.V., 1998.

[19] Rosenberg, A., J. DeLong, C-H. Lee, B-H. Juang, and F. Soong. "The Use of Cohort Normalized Scores for Speaker Verification" *International Conference on Spoken Language Processing in Banff*, University of Alberta, 599 - 602, 1992

[20] J. Friedman, T. Hastie and R. Tibshirani, “Additive logistic regression: a statistical view of boosting”, *Dept. of Statistics, Stanford University Technical Report, 1998.*

A Voice-Centric Multimodal User Authentication System for Fast and Convenient Physical Access Control

Philippe R. Morin and Jean-Claude Junqua

Panasonic Speech Technology Laboratory of Panasonic Technologies Company
A Division of Matsushita Electric Corporation of America
3888 State Street, Suite #202, Santa Barbara, CA 93105
{phm, jcj}@research.panasonic.com

Abstract

In this paper, we present a voice-centric multimodal user authentication system called "BioAxs" that was deployed at our facility to provide fast and convenient physical access control to the laboratory. After discussing the convenience and robustness features present in the system, we describe the core components of the speaker verification engine that revolves around a real-time passphrase spotting strategy. Finally, we describe the multimodal authentication procedure and conclude with experimental field results obtained over a period of 14 months.

1. Introduction

The BioAxs Project initially started as a research and development framework for the study and improvement of PSTL's core speaker verification technology under real conditions. One of the major requirements for this study was the ability to collect and process a large amount of data under real conditions. For that reason, the task of physical access control that could be used by all employees on a daily basis for entering the building was selected. Based on the initial feedback from a first prototype, the approach rapidly evolved into a multimodal user authentication system with a mandate centered on user convenience and robustness. Currently, the system that is located at the building's main entrance door services an average of 140 authentication requests per day for about 35 enrolled users.

2. Mandate and system overview

During the early stage, it became apparent that, in the context of our task, a fast and robust interaction model was a necessity. Convenience became therefore a primary concern for success since employees could always resort to using their key to enter the building. An access terminal housing two biometric modalities (fingerprint and voiceprint) and one non-biometric modality (keypad) was built and installed outside near the main entrance door (first prototype deployed in April 2002). Figure 1 shows a picture of the actual biometric terminal. The terminal is connected to a desktop computer located inside the building via a USB connection.

The access terminal can run in monitoring mode or in user mode. In monitoring mode, the terminal monitors the three sensors in parallel to provide multimodal access control. As explained later in more detail, the authentication procedure enables single modality user authentication for fast interaction, and multi-modality is used to provide smooth uncertainty recovery. In user mode, the terminal allows users to manage their account and to run commands. In that mode, users must first login by entering their 10-digit account number. Once recognized, authorized users can, for instance, enroll (or re-enroll) their voiceprint as well as adapt their existing voiceprint. This self-service mode does not require the need for a system administrator. The system is available to all employees and to a selected number of frequent visitors (e.g. employees of United Postal Service).

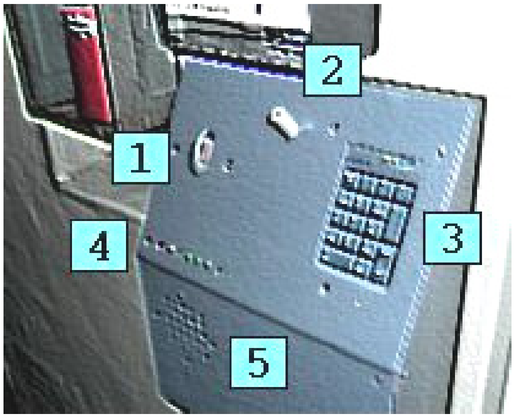


Figure 1. Picture of the biometric terminal showing (1) the fingerprint scanner, (2) the microphone, (3) the keypad, (4) the LED rack, and (5) the loudspeaker components.

3. Overview of the user convenience and robustness features

Convenience, performance, and robustness are primary concerns for global acceptance in real-world applications.

In that respect, a multimodal strategy is specifically advantageous to deal with:

- User preferences: Some users may dislike a given modality or may feel uncomfortable in providing biometric samples for it.
- Disabilities: Some users may not be able to interact with all the modalities due to physical or mental disabilities.
- Redundancy: The environment may render some modalities unusable (e.g. loud noise in the case of voiceprint verification) or the user may be temporarily impaired (e.g. dirty or cut finger in the case of fingerprint verification).
- Verification uncertainties: All modalities have limitations that can be characterized by their respective False Acceptance Rate (FAR) and False Rejection Rate (FRR) distributions. By combining multiple modalities together, verification uncertainties can be virtually eliminated.

Modalities can be categorized based on their activation requirements. Modalities such as fingerprint and keypad require contact and inherently bundle the activation and verification phases into a single step. On the other hand, modalities such as speaker and face verification only require proximity. The proximity paradigm can offer maximum user convenience 1) when the modality does not require some type of external activation and 2) when the proximity constraints are not too restrictive. To provide users with a fast and convenient interaction model, a speaker verification engine was therefore developed based on the following main features:

- Contact-less activation: The system monitors the audio channel continuously without the need for explicit activation such as a push-to-talk button for instance.
- Far-talking microphone: Users can either speak while standing by the biometric box or, more conveniently, they can speak to it as they are approaching; the typical operating range is between 1 and 10 feet.
- Password-dependent voiceprint modeling: Users can register the voice passphrase of their choice to enter the building. The passphrase is used as an active trigger mechanism that allows people (including registered users) to maintain normal conversations in the vicinity on the box.
- Password-spotting input mode: Because the biometric box is located outside the building and is equipped with a far-talking microphone, an input strategy based on automatic endpoint detection was found unreliable in coping with extraneous noises (e.g. street, air conditioning equipment) and babble noise. A spotting strategy is not affected by endpoint errors. User convenience is therefore increased at the expense of an additional burden on the acceptance/rejection module especially in the case of short passwords (e.g. “California”).
- Robust speech front-end: A speech front-end based on sub-band analysis [1] was developed. It incorporates normalization techniques 1) to deal with stationary noises via dynamic spectral band weighting and 2) to increase the robustness with respect to the distance to the microphone via a short-term spectral energy normalization algorithm.

4. Overview of the speech front-end

The speech front-end is based on a sub-band analysis module that generates a power spectrum from the audio stream sampled at 8KHz and filtered with a pre-emphasis coefficient of 0.98. A spectral vector composed of M equally distributed frequency bands is computed every 20 milliseconds (32 bands are typically used for a frequency resolution of 125Hz). Each frequency band is then rescaled using eighth root compression. The front-end only generates static features. Experiments showed that a better accuracy could be obtained with 32 static parameters rather than 16 static and 16 dynamic parameters. The front-end also estimates the average background noise using a decaying average procedure that provides a frequency weighting factor $w_i[t]$ for each band i and instant t. The algorithm uses energy-dependent forgetting factors in order to limit the influence of speech in the estimation process.

To preserve and exploit the redundant information and specificity of the voice contained in the audio signal no cepstral transformation, de-correlation or dimension reduction is performed. In contrast, other analyses such as fifth order PLP [2] tend to purposefully discard some of the speaker’s characteristics.

The frequency band weights mentioned earlier are used at matching time during local distance computation to minimize the impact of the ambient noise. Every 20 milliseconds, energy normalization is performed over a Time Spectral Pattern (TSP) of 300 milliseconds by computing a local loudness factor $E[t]$ as follows:

$$E[t] = \sqrt{\sum_{i=1}^{M} w_i[t] \cdot \sum_{j=-N}^{+N} x_i[t+j]^2 \Big/ (2N+1) \cdot \sum_{i=1}^{M} w_i[t]}$$

where $x_i[t]$ and $w_i[t]$ represents, respectively, the compressed spectral value and the weighting factor at instant t for band i. The final energy-independent parameter vector $y[t]$ is then computed as follows:

$$y_i[t] = \left(x_i[t] - E[t]\right) / E[t]$$

Figure 2 shows a 3D graphical view of the resulting analysis for the phrase “Beautiful Day”. The front-end provides the $w[]$, $x[]$ and $y[]$ streams to the enrollment and verification procedures. The *x[]* stream which preserves the original loudness information of the speech is solely used for the purpose of automatic endpoint detection at training time.

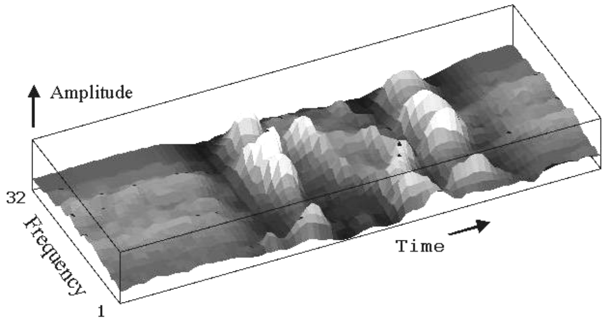


Figure 2. 3D view of the speech analysis *y[t]* generated by the front-end module for the phrase "Beautiful Day" that shows the evolution of the normalized frequency spectrum (32 bands) over time.

5. Text-dependent voiceprint modeling

The enrollment of a new speaker requires five repetitions of a user-selected passphrase and resembles the training procedure of a speaker-dependent word model. Table 1 shows some of the typical passphrases enrolled with the system. It is important to notice that most users prefer to use short passphrases.

Table 1. List of typical user-selected passphrases

California	Garden View
7423	Smoking Gun
Osaka Japan	Copenhagen
Beautiful Day	West Virginia
Geronimo	Treasure Island

A voiceprint model is built by first finding the central repetition using a Dynamic Time Warping (DTW) alignment algorithm. The central repetition is the repetition for which the average alignment distance with respect to all the other repetitions is minimum. The model is then computed by aligning each repetition with the central repetition, and by averaging the sets of aligned parameter vectors. The interactive enrollment procedure verifies however the integrity of the repetitions at each step by building temporary voiceprint models. If the alignment score of a new repetition with respect to the partial model is too low, it is then discarded and the user is re-prompted. If the mismatch occurs on the second repetition, the enrollment procedure is restarted from the beginning.

A voiceprint adaptation function is also available to improve the robustness of the model with speech data from separate sessions. That function must currently be initiated by the users themselves (i.e. via supervised adaptation). Actual statistics show that about 35% of the enrolled users are using that feature.

6. Voiceprint detection and verification

The detection and verification tasks are performed in parallel and in real-time. Figure 3 summarizes the general process that relies on the following three core modules:

1) The *Measurer* module performs frame level matches and provides two types of local similarity scores,
2) The *Aligner* module performs template level matches for all active voiceprint models,
3) The *Spotter* module monitors template score trajectory curves in order to accept or reject hypotheses.

As detailed below, the *Aligner* uses the analysis stream to generate passphrase-dependent data streams which are in turn monitored by the *Spotter*.

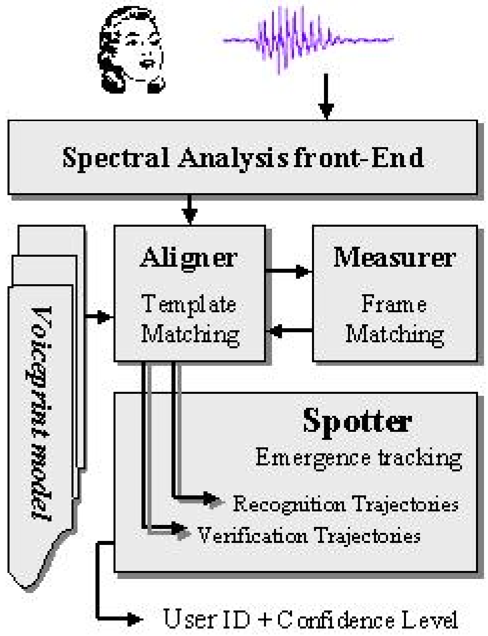


Figure 3. Block diagram of the detection and verification process.

6.1. The Measurer module

To estimate the degree of similarity between a model frame M_i^j (j^{th} parameter frame of voiceprint model *i*) and a test frame *T*, the *Measurer* computes 1) a frame recognition score S_r and 2) a frame verification score S_v. The biometric information about the speaker is contained in both scores. The recognition score is computed as the Euclidian distance between the two frames and is weighted by the dynamic frequency band weights as follows:

$$S_r\left[M_j^i,T\right]=\sqrt{\sum_{k=1}^{M} w_k\cdot\left(M_j^i[k]-T[k]\right)^2}$$

The recognition score is used for template detection by the *Aligner*. On the other hand, the verification score measures more specifically the degree of similarity to a specific voice and is computed as follows:

$$S_v[M_j^i,T]=\frac{S_r[*,T]-S_r[M_j^i,T]}{S_r[*,T]+S_r[M_j^i,T]}$$

where $S_r[*,T]$ represents the background recognition score of test frame *T*. It is estimated by first matching the test frame *T* against all the parameter frames of all voiceprint models and is determined as follows:

$$S_r[*,T]=\mu_T-\omega\cdot\sigma_T$$

where μ_T and σ_T respectively represent the mean and standard deviation of the recognition score distribution and ω is a control coefficient which was experimentally adjusted to 1.4. The verification score has a value ranging from -1.0 (high dissimilarity) to +1.0 (high similarity). The verification score tends to map non speaker-specific frames to the neutral value 0.0.

6.2. The Aligner module

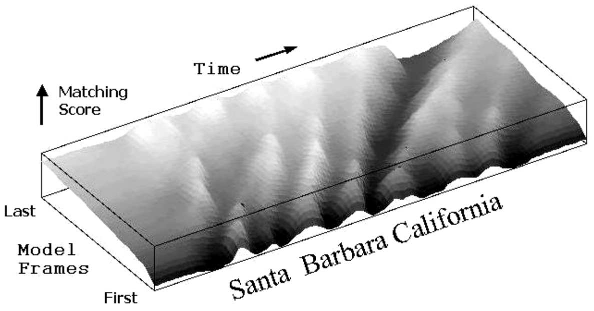


Figure 4. 3D view of the endpoint-free alignment process between a text-dependent model of the passphrase "California" and the test utterance "Santa Barbara California".

The *Aligner* performs matches at the template level. Pattern matching is performed continuously and in parallel for all active voiceprint template models. The *Aligner* computes, at each instant *t* and for each template model, a set of scores that correspond to the best alignment of the template when constrained to end at instant *t*. More specifically, a template recognition score and a template verification score are generated using a DTW algorithm. The search procedure uses a speech recognition criterion exclusively (i.e. speech recognition is preferred over speaker verification during the alignment). Alignment penalties are used to account for insertions and deletions when computing template recognition scores. Template verification scores are computed in parallel with the recognition-centric search but alignment penalties are not used in this case.

Figure 4 shows a 3D view illustrating the endpoint-free alignment process. The valley-shaped depression crossing the template model shows the portion of the signal where a good match for passphrase "California" is measured.

6.3. The Spotter module

The *Spotter* is responsible for the actual detection of enrolled voiceprint templates. This task is achieved by monitoring the voiceprint-dependent score trajectory curves generated by the *Aligner*. These curves typically exhibit a stable area which corresponds to the alignment of the templates against the background noise (i.e. when pattern has not been spoken). That stable value is template-dependent and referred to as "idle score" later. It is continuously estimated and updated by delayed decaying average to account for template duration.

The spotting strategy is based on the principle of relative emergence. The *Spotter* measures the global emergence of each voiceprint model at each instant. The global emergence is computed as a combination of 1) the recognition emergence and 2) the verification emergence. If the global emergence exceeds a given threshold, a detection event is sent to the application, otherwise silent rejection is performed. The emergence of trajectory *S[t]* at instant *t* is defined as:

$$Em[t]=\left(f(S[t])/f(\hat{S}[t])\right)^{\alpha}$$

where $\hat{S}[t]$ represents an estimate of the trajectory's idle score, where *f()* is a transformation function that maps scores into distance-like values (the recognition and verification scores are currently remapped with affine transformations), and where α is a compression/expansion exponent currently set to 0.5.

The emergence principle has the property of normalizing the score trajectories. The normalization compensates 1) for the static differences between templates that are due to differences in phonetic content, and 2) for dynamic differences that occur under different noise conditions. Static differences are more specifically explained by the fact that each phoneme in the language (e.g. '*s*', '*zh*', '*ah*') responds differently in terms of mean score distribution when matched to a given stationary background noise. This natural bias is therefore automatically compensated for. Figure 5 shows the global emergence trajectory curves obtained for a male speaker with respect to his own voiceprint template under three different noise conditions. Brown noise was added to the original audio file (measured at 12 dB SNR) to generate the 8 and 3 dB SNR cases. In the example, the *Spotter* could detect and verify the true user's speech at 12 and 8 dB SNR but could not detect it at 3 dB SNR. If voiceprint detection should occur, the shape of the trajectory typically starts with a flat area centered on the neutral value 1.0 (i.e. where only background is matched), continues with a rise (i.e. the passphrase has been partially spoken at that point) and ends with a fall (i.e. the passphrase has been spoken in its entirety). The degree of match is measured by the depth below the 1.0 idle line.

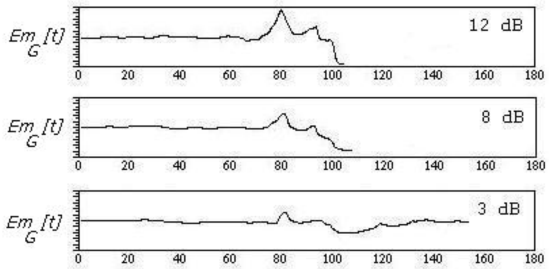


Figure 5. Global emergence versus time *Em$_G$[t]* for the passphrase "California" under three noise level conditions spoken by the true speaker. The detection performance is impacted by the Signal-to-Noise Ratio. At lower SNR where the utterance is being masked by the noise, detection does not occur.

Figure 6 shows, on the other hand, the global emergence trajectory curves for a male imposter (knowing the true user's passphrase) under the same noise conditions.

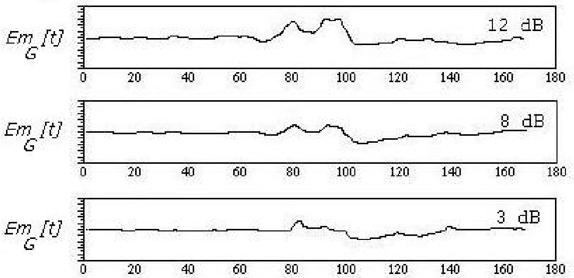


Figure 6. Global emergence versus time *Em$_G$[t]* for the passphrase "California" under three noise level conditions spoken by an imposter. In the example, the observed fall is not deep enough to trigger detection.

The emergence criterion tends to penalize words and phrases that have a higher concentration of fricative and nasal sounds. Words in that category (e.g. 'fishes') can become difficult or virtually impossible to spot due to their higher confusability with the background.

7. The BioAxs authentication procedure

The authentication procedure is primarily unimodal in order to speed up the door access process but all modalities (keypad, fingerprint and voiceprint) are available at all times. Upon successful authentication, the entrance door's contact relay is automatically activated and the name of the verified user is played back along with a series of beeps (from 1 to 5 beeps) indicative of the level of confidence. The multimodal approach helps in the recovery of imperfect matches. An ambiguity occurs when the authentication score is close to the modality's Equal Error Rate. In that case, the security constraints of the helping modality can be reduced without compromising the overall security level, which in the end results in a more robust protocol.

In the case where the user initiates the authentication process by saying his/her voice passphrase, one of three conditions can occur. Based on the authentication score, the system may 1) grant access, 2) deny access or 3) request additional credentials via another modality. In the latter case, the user can either place his/her finger on the scanner or enter his/her "magic" key (currently that key corresponds to first digit of the user's account number) on the keypad. If the credentials are compatible with the hypothesized identity (cross-validation) then access is granted, it is denied otherwise. The user is however still allowed to retry by voice.

Other multimodal strategies could be used. For instance, single modality access could be in effect during core business hours alone. During non-core hours such as at night or during weekends multimodal access (e.g. 2 out of 3 modalities) could be required to enter the building. In that case, the second modality brings increased security at some additional expense in user convenience.

8. Experimental field results

Over a period of about 14 months, the authentication system has serviced an average of 140 authentication requests per day for about 35 enrolled users out of which the vast majority (95%) are voiceprint-initiated requests. The remaining 5% are mostly fingerprint-initiated access requests. Keypad-initiated access, although available, is virtually not used since the process of entering an account number is slow and tedious.

The vast majority of users are very pleased with the convenience brought by the system and keys are very rarely used. The speaker verification module's performance measured under these real-environment conditions is about 8% False Rejection Rate for 0.1% False Acceptance Rate with 2.8% Equal Error Rate. About 37% of these initial rejections are however recovered via multi-modality (i.e. voiceprint or keypad cross-validation) reducing the False Rejection Rate to about 5%.

All the speech data is recorded on disk for database collection purposes. Manual examination of some the audio files clearly indicates that intra-speaker variability (e.g. pitch, enunciation clarity, loudness, prosody changes) is not a negligible phenomenon. The natural variability is however difficult to measure. The phenomenon is exacerbated by the fact that users have the tendency at times to only achieve the articulation level needed to pass the authentication test.

The data collected at the entrance door was also used in the context of a Gaussian Mixture Model (GMM) system. GMM systems [3] use a statistical approach based on single-state Hidden Markov Modeling. It has become a popular state-of-the-art approach for text-independent speaker verification and identification tasks. The GMM system developed at our laboratory uses an MFCC front-end [4] generating 32 ceptral parameters (16 static + 16 dynamic parameters) every 10 milliseconds that are computed from 64 Mel-frequency bands. The front-end uses on-line Cepstral Mean Subtraction (CMS) for channel and noise robustness. A Universal Background Model (UBM) consisting of 256 Gaussian components is used. An adaptive procedure generates speaker models of variable size (i.e. from 64 to 96 Gaussian components per GMM model) to compensate for the differences in duration and in variability across the speaker enrollment data sets. The system uses a background-dependent frame dropping procedure to eliminate non-speech data frames at enrollment and verification time.

The performance of the GMM system obtained on the BioAxs data task is slightly better than the template-based approach and was measured at about 7% False Rejection Rate for 0.1% False Acceptance Rate with 2.4% Equal Error Rate. Unlike the BioAxs system, the GMM system does not however provide recognition results (i.e. the inherent time constraints within a passphrase are not preserved by the GMM modeling method) and therefore recognition errors as well as spotting errors are not accounted for.

Table 2. Performance of the GMM system in text-independent mode on the YOHO database as a function of the number of utterance used for verification.

# of utterances	Equal Error Rate
1	1.91 %
2	0.94 %
4	0.55 %

For comparison purposes, the same GMM system was also tested on the YOHO database in text-independent mode. The YOHO database that is available through the Linguistic Data Consortium (LDC) consists of a collection of 3-number combination lock utterances (e.g. “27-51-83”) from 138 speakers. In this case, the UBM size was increased to 512 Gaussian components and speaker models were built with a fixed size (128 Gaussian components per model). Table 2 shows the system’s performance on that database when trained with 50% of the available training data set. The performance of the GMM system on the YOHO database is comparable to that of other systems [5] [6] even though the system was tuned for text-dependent use.

9. Conclusion

We presented a voice-centric multimodal user authentication system called “BioAxs” which has been deployed at our laboratory to provide physical access control. More specifically, we focused our discussion on the architecture and technology choices that were adopted to provide a fast, convenient and robust interaction model. It was shown that multi-modality is a powerful and natural tool to enable both increased usability to users and increased security to resources. The BioAxs system is extensively used by all employees who all like the convenience of using short voice passphrases to enter the building. Although the current performance has reached a satisfactory level under challenging real environment conditions, further investigation on intra-speaker variability and noise robustness is needed. Ultimately the current False Acceptance Rate (i.e. 8% FAR for 0.1% FRR) must be reduced to increase the level of performance in such a way that users do not have to be too careful when talking to the system.

Our current interest focuses on the improvement of the modeling framework to make use of more statistical information in conjunction with unsupervised adaptation techniques.

10. References

[1] David Kryze, Luca Rigazio, Ted Applebaum and Jean-Claude Junqua, *"A new noise-robust subband front-end and its comparison to PLP"*. The 1999 International Workshop on Automatic Speech Recognition and Understanding, December 12-15, 1999, Keystone, Colorado, USA

[2] Hynek Hermansky, *“Perceptual linear predictive (PLP) analysis of speech”*, JASA 1990 Volume 87, Number 4, Pages: 1738-1752.

[3] D. A. Reynolds and R. C. Rose, “*Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*”, IEEE Trans. on Speech and Audio Processing, 3(1): 72-83, 1995.

[4] S.B. Davis & P. Mermelstein (1980), "*Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*", IEEE Trans. on ASSP 28, 357-366.

[5] J. P. Campbell, Jr., “*Speaker recognition: A tutorial*”, Proceedings of the IEEE, vol. 85, pp. 1437–1462, Sept. 1997.

[6] William M. Campbell and Khaled T. Assaleh, "*Polynomial classifier techniques for speaker verification*", in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 1999, pp. 321-324**.**

Face Recognition Using 2D and 3D Facial Data

Kyong I. Chang Kevin W. Bowyer Patrick J. Flynn

Computer Science & Engineering Department
University of Notre Dame
Notre Dame, IN 46556
{kchang,kwb,flynn}@cse.nd.edu

Abstract

Results are presented for the largest experimental study to date that investigates the comparison and combination of 2D and 3D face recognition. To our knowledge, this is also the only such study to incorporate significant time lapse between gallery and probe image acquisition, and to look at the effect of depth resolution. Recognition results are obtained in (1) single gallery and a single probe study, and (2) a single gallery and multiple probe study. A total of 275 subjects participated in one or more data acquisition sessions. Results are presented for gallery and probe datasets of 200 subjects imaged in both 2D and 3D, with one to thirteen weeks time lapse between gallery and probe images of a given subject yielding 951 pairs of 2D and 3D images. Using a PCA-based approach tuned separately for 2D and for 3D, we find that 3D outperforms 2D. However, we also find a multi-modal rank-one recognition rate of 98.5% in a single probe study and 98.8% in multi-probe study, which is statistically significantly greater than either 2D or 3D alone.

1. Introduction

The identification of the human face in 2D has been investigated by many researchers, but relatively few 3D face identification studies have been reported[1, 2, 3, 4, 5]. One of the main motivations of 3D face recognition is to overcome the problems in general 2D recognition methods resulting from illumination, expression or pose variations.

This study deals with face recognition using 2D and 3D. Each modality captures different aspects of facial features, 2D intensity representing surface reflectance and 3D depth values representing face shape data. Even though each modality has its own advantages and disadvantages depending on certain circumstances, there is often some expectation that 3D data should yield better performance. However, no rigorous experimental study has been reported to validate this expectation. The experiments reported in this study are aimed at (1) examining the spatial / depth resolution needed for 3D face recognition (2) testing the hypothesis that 3D face data provides better biometric performance than 2D face data, using the PCA-based method, and (3) exploring whether a combination of 2D and 3D face data may provide better performance than either one individually in both a single probe study and a multiple probe study.

This is an extension of our earlier work [6]. We have expanded the size of the dataset and have improved the method of geometric normalization used in the 2D and 3D PCA algorithms, resulting in improved recognition performance, both individually and in combination. We have also examined the effect of depth resolution on performance of 3D recognition.

2. Previous Work

In this section, methods that use multiple types of biometric sources for identification purposes, multi-modal biometrics, are reviewed. The term “multi-modal biometrics” is used here to refer to the use of different sensor types without necessarily indicating that different parts of the body are used. The important aspects of these multi-modal studies are summarized in Table 1. Due to the effectiveness of combining multiple biometrics, such studies are included as well to review their data fusion methods, types of biometric sources and the size of experimental dataset. In addition to recognition methods based solely on the human face, there are other recognition methods using multiple biometric sources in addition to face data. One commonality of the studies described in Table 1 is that identification based on multiple sensors / biometrics sources provides overall performance improvement.

3. Methods and Materials

3.1. 2D and 3D Face Recognition Using PCA

Extensive work has been done on face recognition algorithms based on PCA, popularly known as “eigenfaces” [20]. A standard implementation of the PCA-based algorithm [21] is used in the experiments reported here.

Table 1: Multi-biometrics studies for personal identification

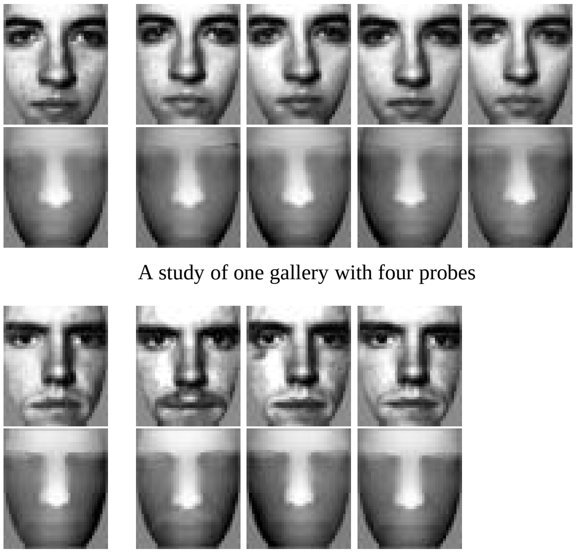
Source (year)	Biometric sources	Fusion methods	Set size
Wang ('03) [7]	Face, Iris	metric-based	90
Chang ('03) [8]	Face, Ear	pixel-based	111
Shakhnarovich('02) [9]	Face, Gait	metric-based	26
Ross ('01) [10]	Face, Hands Fingerprint	metric-based	50
Frischholz ('00) [11]	Face, Voice, Lip Movement	metric-based	150
Ben-Yacoub ('99) [12]	Face, Voice	metric-based	37
Hong ('98) [13]	Face, Fingerprint	metric-based	64
Bigun ('97) [14]	Face, Voice	metric-based	40
Kittler ('97) [15]	Face, Profile Voice	metric-based	37
Brunelli ('95) [16]	Face, Voice	metric- / rank-based	89
Studies that integrate multiple types of *facial data*			
Chang ('03) [6]	2D frontal& 3D shape	metric-based	278
Wang ('02) [17]	2D frontal & 3D shape	metric-based	50
Beumier ('00) [18]	2D frontal & 3D profiles	metric-based	120
Achermann ('96) [19]	2D frontal& 2D profile	metric-/ rank-based	30

3.2. Normalization

The main objective of the normalization process is to minimize the uncontrolled variations that occur during the acquisition process and to maintain the variations observed in facial feature differences between individuals. The normalized images are masked to omit the background and leave only the face region (see Figure 1). While each subject is asked to gaze at the camera during the acquisition, it is inevitable to obtain data with some level of pose variations between acquisition sessions.

The 2D image data is typically treated as having pose variation only around the *Z* axis, the optical axis. The PCA software [21] uses two landmark points (the eye locations) for geometric normalization to correct for rotation, scale, and position of the face for 2D matching. However, the face is a 3D object, and if 3D data is acquired there is the opportunity to correct for pose variation around the *X*, *Y*, and *Z* axes.

A transformation matrix is first computed based on the surface normal angle difference in *X* (roll) and *Y* (pitch) between manually selected landmark points (two eye tips and center of lower chin) and predefined reference points of a standard face pose and location. Pose variation around the *Z* axis (yaw) is corrected by measuring the angle difference between the line across the two eye points and a horizontal line. At the end of the pose normalization, the nose tip of every subject is transformed to the same point in 3D relative to the sensor (see Figure 2). The geometric normalization in 2D gives the same pixel distance between eye locations to all faces. This is necessary because the absolute scale of the face is unknown in 2D. However, this is not the case with a 3D face image, and so the eye locations may naturally be at different pixel locations in depth images of different faces. Thus, the geometric scaling was not imposed to 3D data points as it was in 2D. We found that missing data problems with fully pose-corrected 2D outweighed the gains from the additional pose correction [6], and so we use the typical Z-rotation corrected 2D. Problems with the 3D are alleviated to some degree by preprocessing the 3D data to fill in holes and remove spikes (see Figure 3). This is done by median filtering followed by linear interpolation using valid data points around a hole.



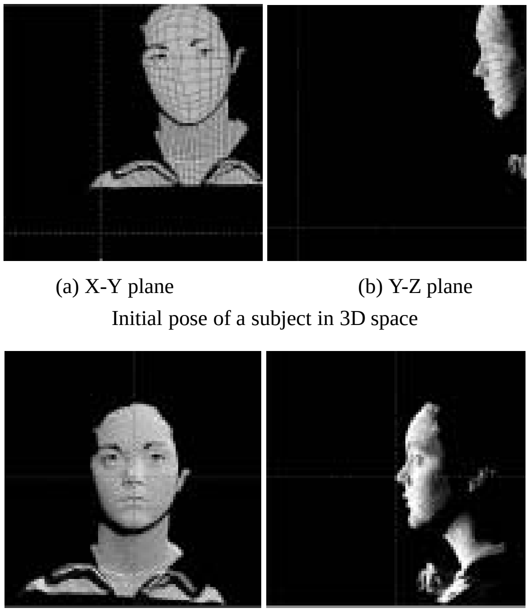
A study of one gallery with four probes

A study of one gallery with three probes

Figure 1: Examples of masked images in 2D and 3D

3.3. Data Collection

A gallery image is an image that is enrolled into the system to be identified. A probe image is a test image to be



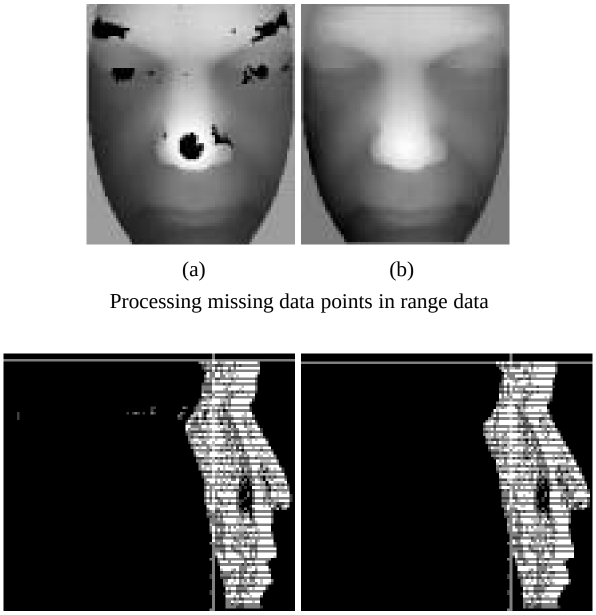
(a) X-Y plane (b) Y-Z plane

Initial pose of a subject in 3D space

(a) X-Y plane (b) Y-Z plane

Corrected pose of a subject in 3D space

Figure 2: Pose normalization



(a) (b)

Processing missing data points in range data

(c) (d)

Processing spike noise in range data

Figure 3: Preprocessing in 3D data points

matched against the gallery images. Images were acquired at the University of Notre Dame between January and May 2003. Two four-week sessions were conducted for data collection, approximately six weeks apart. The first session is to collect gallery images and the second session is to collect probe images for a single probe study in mind. For a study with multiple probes, an image acquired in the first week is used as a gallery and images acquired in later weeks are used as probes. Thus, in the single probe study, there are at least six and as many as thirteen weeks time lapse between the acquisition of gallery image and its probe image, and at least one and as many as thirteen weeks time lapse between the gallery and the probe in the multiple probe study. All subjects completed an IRB-approved consent form prior to participating in each data acquisition session. A total of 275 different subjects participated in one or more data acquisition sessions. Among 275 subjects, 200 participated in both a gallery acquisition and a probe acquisition. Thus, there are 200 individuals in the single probe set, the same 200 individuals in the gallery, and 275 individuals in the training set. The training set contains the 200 gallery images plus an additional 75 for subjects whom good data was not acquired in both the gallery and probe sessions. And for the multiple probe study, 476 new probes are added to the 200 probes, yielding 676 probes in total. The training set of 275 subjects is the same as the set used in the single probe study.

In each acquisition session, subjects were imaged using a Minolta Vivid 900 range scanner. Subjects stood approximately 1.5 meter from the camera, against a plain gray background, with one front-above-center spotlight lighting their face, and were asked to have a normal facial expression ("F_A" in FERET terminology [22]) and to look directly at the camera. Almost all images were taken using the Minolta's "Medium" lens and a small number of images was taken with its "Tele" lens. The height of the Minolta Vivid scanner was adjusted to the approximate height of the subject's face, if needed. The Minolta Vivid 900 uses a projected light stripe to acquire triangulation-based range data. It also captures a color image near-simultaneously with the range data capture. The result is a 640 by 480 sampling of range data and a registered 640 by 480 color image.

3.4. Distance Metrics

2D data represents a face by intensity variation whereas 3D data represents a face by shape variation. It is obvious that the "face space" could be very different between modalities. Thus, during the decision process, certain metrics might perform better in one space than in the other. In this experiment, the Mahalanobis distance metric was explored during the decision process for the gallery matching [23].

3.5. Data Fusion

The pixel level provides perhaps the simplest approach to combining the information from multiple image-based biometrics. The images can simply be concatenated together to form one larger aggregate 2D-plus-3D face image. Metric level fusion combines the match distances that are found in the individual spaces. Having distance metrics from two or more different spaces, a rule for combination of combine the distances across the different biometrics for each person in the gallery can be applied. The ranks can then be determined based on the combined distances.

One of the early tasks in data fusion is to normalize the scores that result from the metric function. Scores from each space need to be normalized to be comparable. There are several ways of transforming the scores including linear, logarithm, exponential and logistic [19]. The scores from different modalities are normalized so that the distribution and the range are mapped to the same unit interval.

There are many ways of combining different metrics to achieve the best decision process, including majority vote, sum rule, multiplication rule, median rule, min rule, average rule and so on. Depending on the task, a certain combination rule might be better than others. It is known that the sum rule and multiplication rule generally provide plausible results [24, 19, 9, 7, 6, 18].

In our study, a weight is estimated based on the distribution of the top three ranks in each space. The motivation is that a larger distance between first- and second-ranked matches implies greater certainty that the first-ranked match is correct. The level of the certainty can be considered as a weight representing the certainty. The weight can be applied to each metric as the combination rules are applied. The multi-modal decision is made as follows. First the 2D probe is matched against the 2D gallery, and the 3D probe against the 3D gallery. This gives a set of N distances in the 2D face space and another set of N distances in the 3D face space, where N is the size of the gallery. A plain sum-of-distances rule would sum the 2D and 3D distances for each gallery subject and select the gallery subject with the smallest sum. We use a confidence-weighted variation of the sum-of-distances rule. For each of 2D and 3D, a "confidence" is computed using the three distances in top ranks as *(second distance - first distance) / (third distance - first distance)*. If the difference between the first and second match is large compared to the typical distance, then this confidence value will be large. The confidence values are used as weights in distance metric. A simple product-of-distances rule produced similar combination results, and a min-distance rule produced slightly worse combination results.

4. Experiments

There are three main parts to this study. The first part is to examine how the recognition performance is affected by the X–Y in both 2D and 3D and depth resolution in 3D data. The second part is to evaluate the performance of 2D and 3D independently in both single and multiple probe studies. Data fusion is considered, in the third part, to combine results at the metric level with different fusion strategies.

The eigenvectors for each face space are tuned by dropping the first *M* and last *N* eigenvectors to obtain an optimum set of eigenvectors. Thus, in general we expect to have a different set of eigenvectors 2D face space versus representing 3D face space. The cumulative match characteristic (CMC) curve is generated to present the results.

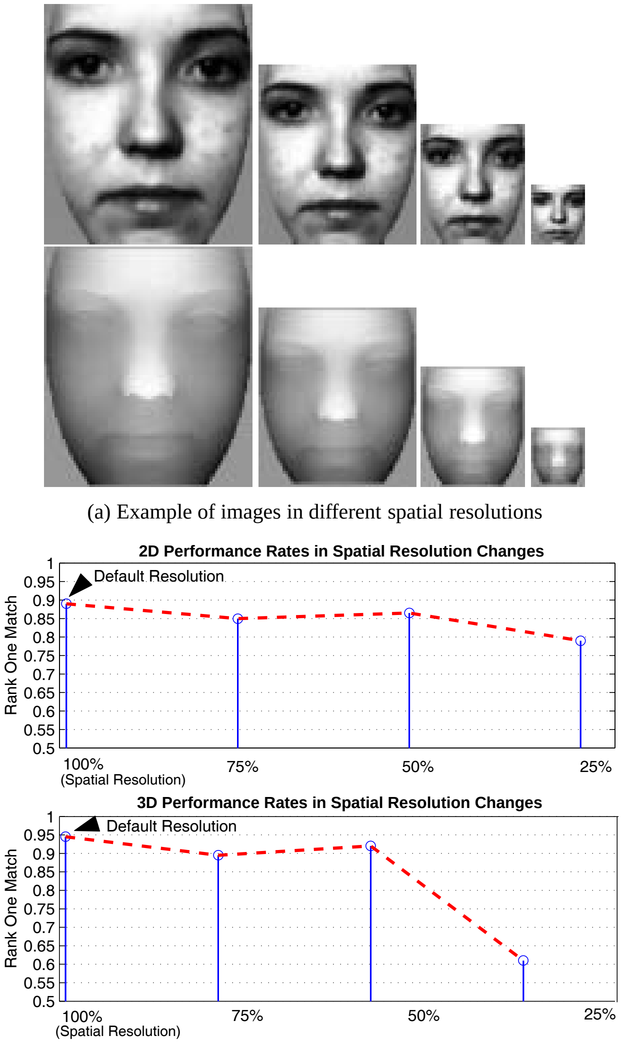
4.1. Experimental Results: X–Y resolution

This experiment looks at the performance rate changes while the spatial resolution is varied in texture and shape images. One average pixel in X axis produced by the Minolta Vivid 900 covers 0.9765mm and one pixel in Y axis covers 0.9791mm of surface area. A typical template size that we initially used was 130 x 150 pixels (a face coverage area of approximately 12.7cm x 14.7cm). Figure 4-(a) shows example of both 2D (top row) and 3D (bottom row) images used for this experiment, starting from the right most, 25%, 50%, 75%, 100% of the original dimension. Thus, every pixel is retrieved in the step of 3.97mm, 1.96mm, 1.31mm and 0.98mm from the original X and Y data points in each image set.

The performance results are shown in Figure 4-(b). The graph is plotted using the first rank match performance rate. Both performance curves begin to drop at the resolution of 1.31mm in X–Y, (in 2D, 89.0% to 85.0%, and 94.5% to 89.5% in 3D). However, the spatial resolution changes attempted in both 2D and 3D suggest that there is no significant difference in performance rates from the original resolution. We believe that performance degradation results from undersampling the face and missing differentiating features. The stiff performance drop has been shown in between 50% and 25% due to the insufficient facial features to be differentiated between subjects in PCA method.

4.2. Experimental Results: Depth resolution

This experiment has a similar purpose as the previous one. However, this examines the depth resolution required to maintain the performance rate from the original depth resolution. According to the Minolta Vivid 900 specification, its depth accuracy level may be obtained at 0.35mm. One way to vary the original resolution is to change the precision level in floating point values of the Z coordinate. A lower limit on precision could be $10^{-6}mm$. However, the camera-to-subject distance and lens combination used in our acquisition likely support an actual depth resolution of no better than about 0.5mm on average. Fourteen different resolutions were examined so that every pixel value representing the actual coordinate is retrieved in the unit



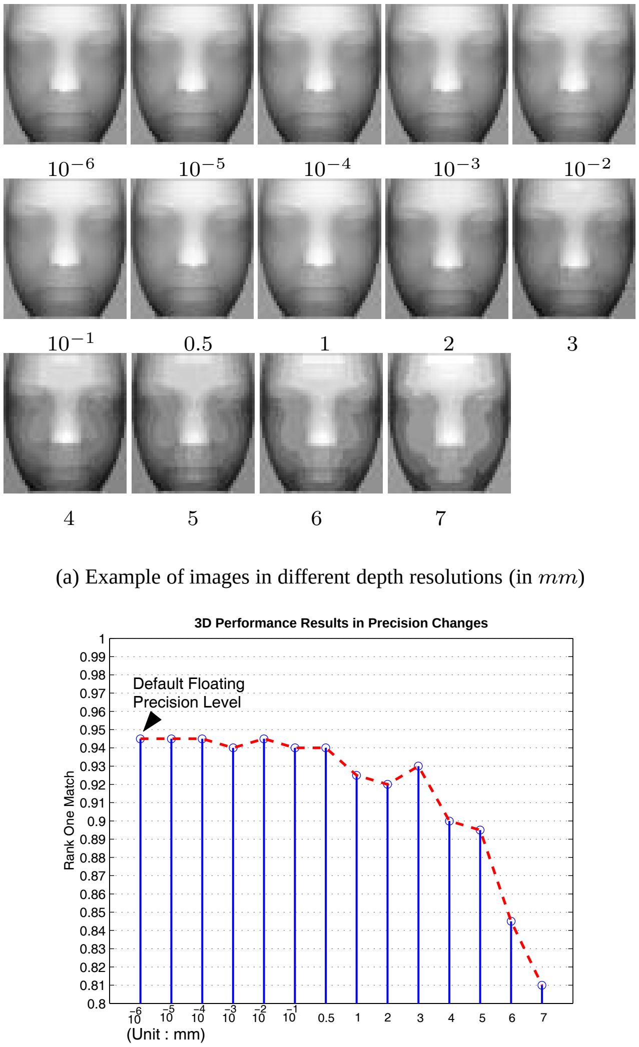
(a) Example of images in different spatial resolutions

(b) Different spatial resolutions

Figure 4: Experiment in spatial resolutions changes

of $10^{-6}mm$, $10^{-5}mm$, $10^{-4}mm$, $10^{-3}mm$, $10^{-2}mm$, $10^{-1}mm$, $0.5mm$, $1mm$, $2mm$, $3mm$, $4mm$, $5mm$, $6mm$ and $7mm$ as shown in Figure 5-(a). As shown in Figure 5-(b), the overall performance rate decreases as the depth resolution gets coarser. It becomes prominent after $3mm$.

However, it is interesting to note that the performance rates between $0.5mm$ and $3mm$ maintain remarkably close to the original resolution (within 2.5%). This may be partially because as the resolution gets coarser, random noise would be suppressed. As it gets even coarser, a face surface becomes overly contoured and identification suffers from such coarsely quantized surfaces.



(a) Example of images in different depth resolutions (in mm)

(b) Performance results in different depth resolutions

Figure 5: Experiment in depth resolution changes

4.3. Experimental Results: 2D versus 3D face - Single probe study

This experiment is to investigate the performance of individual 2D eigenface and 3D eigenface methods, given (1) the use of the same PCA-based algorithm implementation, (2) the same subject pool represented in training, gallery and probe sets, and (3) the controlled variation in one parameter, time of image acquisition, between the gallery and probe images. A similar comparison experiment between 2D and 3D acquired using stereo-based system was also performed by Medioni *et.al.*[25].

There can be many ways of selecting eigenvectors to accomplish the face space creation. In this study, at first, one vector is dropped at a time from the eigenvectors of largest eigenvalues, and the rank-one recognition rate is computed

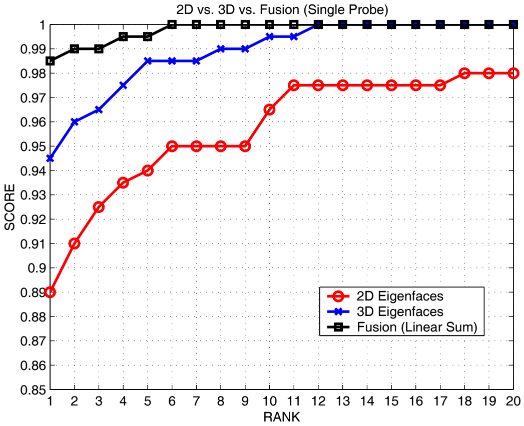


Figure 6: Performance results in single probe study.

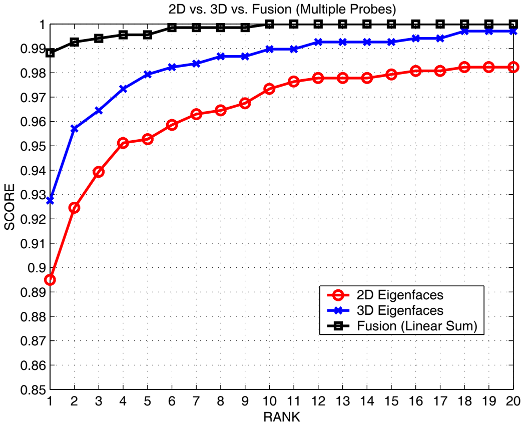


Figure 7: Performance results in multiple probe study.

using the gallery and probe set again each time, and continue until a point is reached where the rank-one recognition rate gets worse rather than better. We denote the number of dropped eigenvectors of largest eigenvalues as *M*. Also, one vector at a time is dropped from the eigenvectors of the smallest eigenvalues, and the rank-one recognition is computed using the gallery and probe set again each time, and continue until a point is reached where the rank-one recognition rate gets worse rather than better. We also denote the number of dropped eigenvectors of smallest eigenvalues as *N*.

During the eigenvector tuning process, the rank-one recognition rate remains basically constant with from one to 20 eigenvectors dropped from the end of the list. This probably means that more eigenvectors can be dropped from the end to create a lower-dimension face space. This would make the overall process simpler and faster. The rank-one recognition rate for dropping some of the first eigenvectors tend to improve at the beginning but it start to decline as *M* gets larger.

After the eigenvectors are tuned, both 2d and 3D are coincided at *M* = 2, and *N* = 20 to create the face spaces. With the given optimal set of eigenvectors in 2D or 3D, the results show that rank-one recognition rate is 89.0% for 2D, and 94.5% for 3D (see Figure 6).

4.4. Experimental Results: Multi-modal biometrics using 2D and 3D

The purpose of this experiment is to investigate the value of a multi-modal biometric using 2D and 3D face images, compared against individual biometrics. The null hypothesis for this experiment is that there is no significant difference in the performance rate between uni-biometrics (2D or 3D alone) and multi-biometrics (both 2D and 3D together). According to Hall [26], a fusion can be usefully done if an individual probability of correct inference is between 50% and 95% with one to seven classifiers. From our results in the previous experiment, it is reasonable to fuse the two individual biometrics which meet this fusion criteria. Figure 6 shows the CMC with the rank-one recognition rate of 98.5% for the multi-modal biometric, achieved by combining modalities at the distance metric level. In the fusion methods that we considered, the multiplication rule showed the most consistent regardless of the particular score transformation. However, the min rule showed lower performance than any other rules in different score transformations (see Figure 8). Also, when the distance metrics were weighted based on the confidence level during the decision process, all the rules result in significantly better performance than the individual biometric. A McNemar's test for significance of the difference in accuracy in the rank-one match between the multi-modal biometric and either the 2D face or the 3D face alone shows that multi-modal performance is significantly greater, at the 0.05 level.

4.5. Experimental Results: 2D face versus 3D face in biometrics - multiple probe study

In these experiments, there will be one or more probes for a subject who appears in the gallery, with each probe being acquired in a different acquisition session separated by a week or more. We are attempting to retrieve more practical use of face identification method by incorporating multiple probes to be matched against the gallery images. The multiple probe dataset consists of 676 probes in total. Subjects might have a different number of probes. For example, there are 200 subjects with 1 or more probes, 166 subjects with 2 or more probes and so on. In the probe dataset, the number of probes can be up to 7 per subject. There might be different rules to determine a correct match given several probes to a gallery. In this experiment, a correct match is

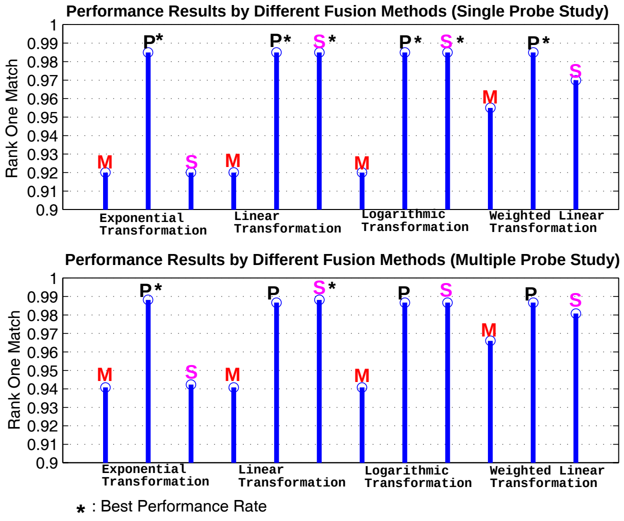


Figure 8: Performance results of fusion schemes used.

measured based on an each individual probe rather than on some function of all probes per subject.

By using the same set of eigenvectors tuned in the single probe study, we achieved similar results as in the previous sections. While 3D performance dropped a little, 92.8%, 2D performance maintains slightly better than the previous experiment, 89.5% (see Figure 7).

After combining these two biometrics in the multiple probes, we also were able to obtain significantly better performance, at 98.8%, than for either 2D or 3D alone. The results of 2D and 3D combination show very similar performance behavior as the single probe study. Product rule performs better than minimum rule regardless of score transformation (see Figure 8). Most combined methods consistently perform significantly better than the single biometrics. A McNemar's test for significance of the difference in accuracy in the rank-one match between the multi-modal biometric and either the 2D face or the 3D face alone shows that multi-modal performance is significantly greater, at the 0.05 level. Thus, significant performance improvement has been accomplished by combining 2D and 3D facial data in both single and multiple probe studies.

5. Summary and Discussion

The value of multi-modal biometrics with 2D intensity and 3D shape of facial data in the context of face recognition is examined in a single probe study and a multiple probe study. This is the largest experimental study (in terms of number of subjects) that we know of to investigate the comparison and combination of 2D and 3D data for face recognition. In our results, each modality of facial data has roughly similar value as an appearance-based biometric. The combination of the face data from both modalities results in statistically significant improvement over either individual biometric. In general, our results appear to support the conclusion that *the path to higher accuracy and robustness in biometrics involves use of multiple biometrics* rather than the best possible sensor and algorithm for a single biometric.

We also have investigated the effect of spatial and depth resolution on recognition performance. This was done by producing successively coarser versions of the original image. The original image has a depth accuracy at $0.35mm$. We found that performance drops only slightly in going to a depth resolution of $0.5mm$, but begins to drop drastically at $4mm$. The pattern of results suggests that it would be interesting to determine a sensor accuracy level needed to meet a specific requirement of face recognition tasks. The accuracy requirement might be vary under different conditions of subjects, such as facial muscle movement, or imaging condition changes. This initial investigation in resolution variation would bring a more explicitly decided resolution level for further experiments.

The overall quality of 3D data collected using a range camera is perhaps not as reliable as 2D intensity data. 3D sensors in the current market are not as mature as 2D sensors. Common problems with typical range finder images include missing data in eyes, cheeks, or forehead as well as several types of noise. These problems would lower the 3D recognition rate in general even though there exist ways of recovering some data in such areas.

The criteria used to decide which combination of eigenvectors to keep is the rank-one recognition rate on the gallery and probe images. So, in a way, the gallery and probe images are used in deciding what eigenvectors to use for the space, and then the results are also reported on the gallery and probe images, thereby "testing on training data". This can be addressed by having a validation set of images to determine the set of eigenvectors to be used during the identification process so that eigenvectors to keep before the performance on the gallery and probe images are obtained.

It is generally accepted that performance estimates for face recognition will be higher when the gallery and probe images are acquired in the same acquisition session, compared to performance when the probe image is acquired after some passage of time [27]. Most envisioned applications for face recognition technology seem to occur in a scenario in which the probe image would be acquired some time after the gallery image. In this context, it is worth noting that the dataset used here incorporates a substantial time lapse between gallery and probe image acquisition.

The dataset used in the experiments reported here will be made available to other research groups as a part of the Human ID databases. See http://www.nd.edu/˜cvrl/ for more information about the dataset and the release agreement.

6. Acknowledgments

This work was supported by the DARPA HumanID program through Office of Naval Research N000140210410 and by the National Science Foundation and the Central Intelligence Agency through NSF under grant EIA 01-30839.

References

[1] L. Lee, B. Moghaddam, H. Pfister, and R. Machiraju, "Silhouette-based 3D face shape recovery," *Graphics Interface*, June 2003.

[2] A. Bronstein, M. Bronstein, and R. Kimmel, "Expression-invariant 3D face recognition," *Audio and Video based Biometric Person Authetication*, pp. 62–69, 2003.

[3] J. Huang, B. Heisele, and V. Blanz, "Component-based face recognition with 3D morphable models," *Audio and Video based Biometric Person Authetication*, pp. 27–34, 2003.

[4] C. Chua, F. Han, and Y. Ho, "3D human face recognition using point signature," *Int'l Conf. on Automatic Face and Gesture Recognition*, pp. 233–238, 2000.

[5] B. Achermann, X. Jiang, and H. Bunke, "Face recognition using range images," in *Proceedings Int'l Conf. on Virtual Systems and MultiMedia '97, Geneva, Switzerland*, Sept. 1997, pp. 129–136.

[6] K. Chang, K. Bowyer, and P. Flynn, "Multi-modal 2D and 3D biometrics for face recognition," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 187–194, October 2003.

[7] Y. Wang, T. Tan, and A. K. Jain, "Combining face and iris biometrics for identity verification," *4th Audio and Video based Person Authentication - Guildford, UK*, 2003.

[8] K. Chang, K. Bowyer, V. Barnabas, and S. Sarkar, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 25, pp. 1160–1165, 2003.

[9] G. Shakhnarovich and T. Darrell, "On Probabilistic Combination of Face and Gait Cues for Identification," *Int'l Conf. on Automatic Face and Gesture Recognition*, pp. 169–174, 2002.

[10] A. Ross and A. Jain, "Information fusion in biometrics," *Audio and Video based Person Authentication*, pp. 354–359, 2001.

[11] R. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," *IEEE Computer*, vol. 33, no. 2, pp. 64–68, Feb. 2000.

[12] S. Ben-Yacoub, "Multi-modal data fusion for person authentication using SVM," *Audio and Video based Person Authentication*, pp. 25–30, 1999.

[13] L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification," *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 20, no. 12, pp. 1295–1307, 1998.

[14] E. Bigun, J. Bigun, B. Duc, and S. Fischer, "Expert conciliation for multi modal person authentication systems by bayesian statistics," *Audio and Video based Person Authentication*, pp. 311–318, 1997.

[15] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez, "Combining evidence in personal identity verification systems," *Pattern Recognition Letters*, , no. 9, pp. 845–852, 1997.

[16] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 17, no. 10, pp. 955–966, 1995.

[17] Y. Wang, C. Chua, and Y. Ho, "Facial feature detection and face recognition from 2D and 3D images," *Pattern Recognition Letters*, vol. 23, pp. 1191–1202, 2002.

[18] C. Beumier and M. Acheroy, "Automatic face verification from 3D and grey level clues," *11th Portuguese Conf. on Pattern Recognition*, Sept. 2000.

[19] B. Achermann and H. Bunke, "Combination of face classifiers for person identification," *Int'l Conf. on Pattern Recognition*, pp. 416–420, 1996.

[20] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[21] R. Beveridge and B. Draper, "Evaluation of face recognition algorithms (release version 4.0)," URL : http://www.cs.colostate.edu/evalfacerec/index.html.

[22] J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 22, no. 10, pp. 1090–1104, October 2000.

[23] W. Yambor, B. Draper, and J. Beveridge, "Analyzing PCA-based face recognition algorithms: Eigenvector selection and distance measures," *2nd Workshop on Empirical Evaluation in Computer Vision, Dublin, Ireland*, July 1 2000.

[24] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 20, no. 3, pp. 226–239, 1998.

[25] G. Medioni and R. Waupotitsch, "Face modeling and recognition in 3-D," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 232–233, 2003.

[26] D. Hall, *Mathematical Techniques in Multisensor Data Fusion*, Artech House, Norwood, MA., 1992.

[27] J. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and M. Bone, "Facial recognition vendor test 2002: Evaluation report," *NISTIR 6965*, URL: http://www.frvt2002.org/FRVT2002/documents.htm.

Compression For Distributed Face Recognition

Navid Serrano[1], Antonio Ortega[1], Shuang Wu[2], Kazunori Okada[3], and Christoph von der Malsburg[4]

[1]Department of Electrical Engineering-Systems
[2]Department of Physics and Astronomy
[4]Department of Computer Science
University of Southern California
{navidser,aortega,shuangwu,malsburg}@usc.edu

[3]Siemens Corporate Research, Inc.
kazunori.okada@scr.siemens.com

Abstract

We investigate the application of a face recognition system in a distributed environment. Images of faces are captured by clients remotely and transmitted to a server for recognition or authentication using a central database. In many distributed scenarios, bandwidth may be limited and transmission of image data may not be feasible. We assume the client does not have processing limitations and can extract and transmit compressed features. In addition, the compressed features can be used as an alternative representation of the faces in the database and thus reduce storage needs. In this paper we explore the impact of feature compression on face recognition performance. Specifically, we consider the Bochum/USC face recognition system and propose an embedded coding scheme for Gabor-based wavelet features extracted from optimally selected landmarks on the face. Our results show that the impact on recognition rates—even at the highest compression rates—is minimal.

1. Introduction

Distributed processing systems are fast becoming an important area of research. The paradigm of centralized computing is changing as mobile devices and sensors with enhanced processing power and multimedia capabilities proliferate. The emergence of distributed processing systems enables more flexible solutions to existing problems and simultaneously poses new challenges.

Applications such as user authentication can now be carried out using a multitude of devices in a wide variety of locales but must account for bandwidth limitations if data is to be transmitted between clients and a server. Each time a user needs to be authenticated remotely, pertinent information (e.g. authentication features) must be sent to a central server for processing. If the communication channel between the client and the server has limited bandwidth, it may be intractable to transmit large amounts of information. This is a considerable limitation for applications that rely on visual information, such as a face recognition system.

The bandwidth problem can be mitigated if the client shoulders the processing burden entirely. However, there are scenarios where this is neither possible nor desirable. In many authentication systems, for example, it is often necessary to store the database centrally in order to add users dynamically and ensure security. In addition, a central database is easier to maintain. Therefore, it is reasonable to distribute the processing load between clients and the server. Because communication is inevitable in distributed processing systems, the flow of information can be facilitated using some form of compression.

In a distributed face recognition system a server maintains a large database of faces centrally and clients submit face information to the server for recognition. It may be advantageous for the client to extract and transmit the features instead of the face image data in order to avoid bandwidth limitations altogether. To further preserve bandwidth, the features themselves could be compressed prior to transmission.

Compression for distributed processing systems is a growing area of research. Applications such as distributed speech recognition [1], distributed image classification [2], and distributed sensor networks [3] have been studied. Yet, compression for distributed face recognition remains unexplored. As intimated earlier, we are interested, in particular, in a scenario where the client has sufficient processing power but limited bandwidth. Furthermore, we assume the server maintains the database centrally.

Based on the above assumptions, the Bochum/USC face recognition system [4] is a good study case. This system can easily be adapted to a distributed environment because it is based on general principles rather than statistical learning. In fact, classification is based on a simple nearest neighbor distance metric. This implies that faces can be easily added to the database without having to dynamically retrain the classifier. The fundamental part of the algorithm involves an elastic bunch graph technique used to locate specific landmarks on the face from which Gabor wavelet features are extracted.

Prior research has shown that in bandwidth limited cases—where compression is inevitable—the decrease in classification accuracy is less when compressing feature vectors as opposed to the signal itself [1]. Therefore, we attempt to exploit the underlying structure of the Gabor wavelet features as a first step. Although the wavelet features are extracted from selected landmarks on the face, they demonstrate some tendencies common to natural images, including energy concentration in the low-frequency sub-bands. Given these observations, we propose a modified bit-plane coding technique similar to the embedded zerotree wavelet coder proposed by Shapiro [5]. The compression scheme also has the advantage of being fully embedded, meaning that finer renditions of the features are transmitted progressively. We show that the embedded coding scheme can achieve a bit-rate of 0.13 bits per pixel (based on 128 x 128 pixel images) while decreasing the overall face recognition rate on average by only 1%. Finally, we show that face recognition performance is considerably better using embedded feature coding compared to image compression using the state of the art JPEG2000 [10] technology, although, obviously, our scheme, unlike JPEG2000, does not provide a decodable version of the full image.

As mentioned earlier, in a distributed scenario the limiting factor is bandwidth. Although this is the main focus of our paper, we also address the case (which is not necessarily distributed) where the limiting factor is storage capacity. When the image database is extremely large and/or storage space is limited, it is often desirable to compress the images in the database. An alternative is to build a database of compressed features. In this paper we also evaluate our proposed feature compression scheme for such a scenario, i.e., as a tool to provide a compressed representation of the database. In this respect our proposed compression scheme performed favorably compared to JPEG2000.

The paper is organized as follows. We first provide an overview of the Bochum/USC face recognition system, then discuss feature vector compression, including our proposed embedded coding technique, and conclude with an evaluation of the impact of compression on face recognition.

2. The Bochum/USC Face Recognition System

As discussed earlier, the performance of the Bochum/USC face recognition system is based largely on the efficacy of a bunch graph matching technique [6] used to optimally locate specific landmarks on the face. Gabor wavelet features are then extracted from these landmarks and classified using a simple similarity measure. In terms of feature compression, it is noteworthy that the algorithm does not rely heavily on the generalization capability of a classifier engine. Otherwise, distortion resulting from compression could displace vectors in feature space and consequently affect recognition performance. Finally, it should also be noted that the Bochum/USC face recognition system achieved the best performance among competing algorithms in a FERET test administered between September 1996 and March 1997 [7].

Each elastic graph has 48 nodes placed at specific landmarks on the face. The face recognition system addresses pose variations but for our present purposes we only consider frontal views of the face, as shown in Figure 1.

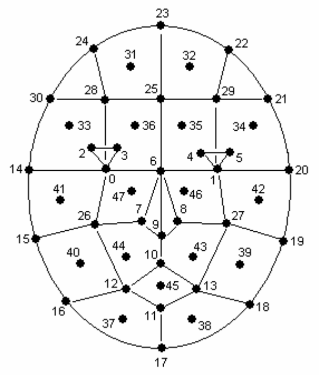


Figure 1. Spatial location of the 48 landmarks on the face.

At each node location, a feature vector, or jet, is extracted using a Gabor-based wavelet expansion:

$$\psi_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{k}^2}{\sigma^2} \exp\left[-\frac{\mathbf{k}^2\mathbf{x}^2}{2\sigma^2}\right]\left[\exp(i\mathbf{kx}) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \quad (1)$$

where $\mathbf{x}$ are the image coordinates, $\mathbf{k} \in \Re^{LD}$ is the wave vector, l=1,…,L is the frequency index, d=1,…,D is the direction index, and $\phi_d = \pi d/D$ is the kernel direction. Let $J_{\mathbf{k}}(\mathbf{x})$ be the Gabor jet extracted at each landmark by convolving the image $I(\mathbf{x})$ with the Gabor wavelet of Eq. (1):

$$J_{\mathbf{k}}(\mathbf{x}) = \int I(\mathbf{x}')\psi_{\mathbf{k}}(\mathbf{x} - \mathbf{x}')d\mathbf{x}' \quad (2)$$

In the Bochum/USC face recognition system, the Gabor wavelet is sampled over L=5 levels in D=8 directions:

$$J_j = a_j e^{i\phi_j} \quad (3)$$

Where a_j and ϕ_j, j=1,…,LD, are the wavelet coefficient magnitude and phase respectively. A model graph is the collection of Gabor jets (i.e. feature vectors) at each of the N=48 landmarks. Therefore, the Gabor jet at each landmark will have 80 elements (2LD) organized in the following manner:

$$\begin{bmatrix} J(\mathbf{x}^1) \\ \vdots \\ J(\mathbf{x}^N) \end{bmatrix} = \begin{bmatrix} a_1^1 & \phi_1^1 & \cdots & a_{LD}^1 & \phi_{LD}^1 \\ \vdots & \vdots & & \vdots & \vdots \\ a_1^N & \phi_1^N & \cdots & a_{LD}^N & \phi_{LD}^N \end{bmatrix} \qquad (4)$$

The similarity between any two jets J and J' can be computed by:

$$S(J,J') = \frac{\sum_j a_j a'_j}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \qquad (5)$$

The similarity between two faces can be obtained by computing the node-wise similarity between corresponding jets on each face, according to Eq. (5). This similarity function is phase insensitive but other similarity measures could be used.

3. Feature Vector Compression

One of the challenges in devising a compression scheme for feature vectors is to identify redundancy, irrelevance, and structure. Assumptions can be made about the features used here. In terms of irrelevance, it is reasonable to assume that certain landmarks impact recognition more than others. In such a case, one option would be to apply a coarser quantization to less important landmarks. Another option is to simply eliminate less important landmarks—equivalent to a bit-rate of 0. In fact, the application of principal component analysis to represent the Gabor wavelet features using fewer parameters was studied in [8].

In terms of redundancy, the symmetry of the facial landmarks could be exploited. For instance, it is reasonable to assume that the Gabor jets extracted from the landmarks on the left eye will be similar to the Gabor jets extracted from the landmarks on the right eye. A cursory study of the images in a database of 800 images showed that the features extracted from the left eye were strongly correlated with features extracted from the right eye. On the other hand, the study also showed that most of the other symmetric landmarks—particularly those near the edge of the face—showed little correlation.

Another option is to exploit the underlying structure of the wavelet coefficients. It has been shown that wavelet coefficients extracted from natural images have a strong energy concentration in the low-frequency sub-bands [9]. This trend can be used favorably for compression, as demonstrated by the use of wavelet coding in the new JPEG2000 image compression standard [10]. However, the Gabor wavelet coefficients are only extracted from selected landmarks. Still, our evaluation of the Gabor jets showed that the wavelet coefficient energy was concentrated in the low-frequency sub-bands and furthermore, the energy of coefficients in high-frequency sub-bands tended to be correlated with the energy in low-frequency sub-bands.

3.1. Embedded Coder

The embedded zerotree wavelet (EZW) algorithm was introduced by Shapiro [5]. The EZW compression scheme codes data in a signal by using efficient entropy coding of its bit-plane representation. Namely, the samples in the signal of interest (e.g. wavelet coefficients) are represented using bit-planes. The embedded nature of the EZW algorithm arises from the fact that each bit-plane is coded separately and transmitted in order of importance from the LSB to the MSB plane. Quantization is achieved by stopping this refinement process, i.e., by not sending bit-planes below a certain threshold. Shapiro proposed using embedded coding on wavelet coefficients given that there is often a correlation between energy values across frequency levels.

The bit-plane coding is accomplished by assigning one of four labels to the wavelet coefficients. If a coefficient is above a threshold T (half the value of the bit-plane), it is assigned a *significant positive* or *significant negative* label, depending on its sign. If the wavelet coefficient is below the threshold and all of its descendants (i.e. the coefficients in higher frequency sub-bands) are also below the threshold, then it is assigned a *zerotree root* label. Otherwise, it is assigned an *isolated zero* label. This means that only four labels are needed to code the coefficients at each bit-plane. Furthermore, if a coefficient is a zerotree root, it does not have to be transmitted at higher bit-planes. The reconstruction value for significant coefficients at each bit-plane is simply $3T/2$

EZW also includes a refinement stage where the reconstructed value of significant coefficients is refined. The refinement is simply $\pm T/4$ and only requires one bit for transmission. Thus, at each bit-plane two bit-streams are transmitted: the significant coefficient bit-stream s and the refinement bit-stream r. Assuming the wavelet coefficients have low-frequency energy compaction, this technique can yield significant savings. EZW is an embedded coder in that the bit-stream is generated and transmitted in order of importance.

Ordinarily, in an image compression scheme, the wavelet coefficients might have to be stored in memory in order to define the tree structure. This is because a typical wavelet transform algorithm generates all the coefficients of each wavelet subband but the tree is formed by grouping coefficients from different bands together. Depending on the size of the original image, this can lead to a considerable memory requirement. However, because in this case the Gabor wavelet coefficients are self-contained in relatively low-dimensional feature vectors, a tree structure (i.e. an energy coefficient dependence) can be easily defined without memory constraints. It should be noted that other embedded coders have been developed since EZW and could be applied here but we choose EZW because of its simplicity.

Note that, while in standard image coding applications, wavelet coefficients for each subband are generated for the whole image, in the Gabor jet case the data that is generated is naturally localized, since the transformation is only done in the neighborhood of the face landmarks. In short, because in our problem we have to operate with localized trees of coefficients at various frequencies it is more natural to extend the EZW tree, rather than group coefficients corresponding to the same frequency and orientation from different landmarks, before proceeding to a JPEG2000-style encoding. Moreover, when localization is preserved in the coding it is possible to tailor the level of quantization to the relative importance for recognition of the different landmarks.

3.2. Modified Embedded Coder for Gabor Jets

We modify the EZW principle to apply it to the Gabor jets extracted using the Bochum/USC face recognition system. The coefficients in the Gabor Jets are infinite precision. Therefore, in order to apply bit-plane coding, they were scaled to 8-bit finite precision. The scaling can be done by normalizing relative to the maximum component of each Gabor jet and then sending this value as side information. Alternatively, the scaling can be done using a global scale factor for each landmark. There is no reason why the coefficients could not be scaled to higher bit representations. The only change would be the number of bit-planes to code.

The first modification to the EZW algorithm is the elimination of the significant negative label since the Gabor jet coefficients are all positive. Another modification is the establishment of parent-child relationships and a scanning order to determine zerotree roots. In natural images, a parent is a wavelet coefficient at any sub-band. Children are coefficients in higher frequency subbands corresponding to the same spatial location in the original image. The scanning is done in zig-zag fashion, according to the traditional dyadic pyramid representation. In the case of the Gabor jets, each of the eight directions is coded separately. We establish parent-child relationships in each direction. The scanning order is from the low-frequency coefficients (l=1) toward the highest frequency coefficients (l=5), as shown in Figure 2. Define the threshold T_i at bit-plane i as:

$$T_i = \begin{cases} 2^{\lfloor \log_2 A \rfloor}, & i = 0 \\ \frac{1}{2} T_{i-1}, & i = 1,...,7 \end{cases}, \tag{6}$$

Where, A is the maximum wavelet coefficient magnitude:

$$A = \max_j a_j \,, \tag{7}$$

The entry in the bit-stream at bit-plane i for a given wavelet coefficient $a_{l,d}$ is assigned one of three values:

$$s_i = \begin{cases} 0, & a_{l,d} \geq T_i \\ 1, & a_{l+j,d} < T_i, \forall j = 0,..., L \\ 2, & a_{l,d} < T_i \end{cases}, \tag{8}$$

Where, l=1,…,L is the decomposition level, d=1,…,D is the direction, and the symbols 0, 1, and 2 represent the significant positive, zerotree root, and isolated zero coefficients respectively. The entry in the refinement bit-stream is simply:

$$r_i = \begin{cases} 0, & T_i \leq a_{l,d} < \frac{3}{2} T_i \\ 1, & \frac{3}{2} T_i \leq a_{l,d} < 2T_i \end{cases}, \tag{9}$$

Example Gabor wavelet coefficients extracted from landmark 17 are shown in Figure 2 (averaged over all images in the database of faces). The decreasing energy trend from low frequency to high frequency coefficients can also be seen in Figure 2.

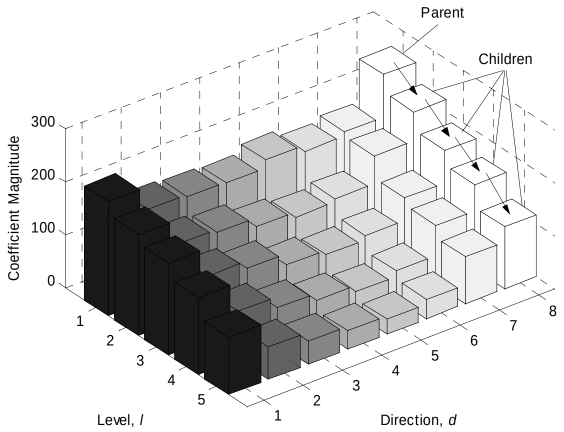


Figure 2. Parent-child dependency and scanning order.

4. Evaluation

We evaluated the proposed compression scheme on a database of 800 faces (frontal view). For each facial image there are 48 Gabor jets. Each jet is coded separately and weighted equally. The embedded coder can be applied to both the magnitude and phase elements but we only consider the magnitude here since the similarity measure of Eq. 5 is phase insensitive. For each bit-plane, the coefficients are assigned one of three labels, zerotrees are established and the entropy of the bit-stream is computed. Since the bit-stream is embedded, the entropy of the bit-stream at any bit-plane b is equal to the sum of entropies of itself and previous bit-planes:

$$H = \sum_{i=0}^{b} \left[\frac{S_i \sum_{j=0}^{2} P(s_i = j) \log P(s_i = j) + R_i \sum_{j=0}^{1} P(r_i = j) \log P(r_i = j)}{S_i + R_i} \right] \quad (10)$$

Here, s_i and r_i are the significant and refinement bit-streams at bit-plane i, respectively, and S_i and R_i are the lengths of s_i and r_i. For an entire model graph, the entropy of each Gabor jet was averaged. The compression ratio *CR* for the embedded coder is simply:

$$CR = \frac{\sum_{k=0}^{255} P(a_j = k) \log P(a_j = k)}{H} \quad (11)$$

Where, $P(a_j = k)$ is the probability that the Gabor coefficient magnitude a_j, j=1,…,*LD*, is equal to the 8-bit value *k*. Hence, Eq. (11), measures the compression ratio of the coefficient magnitudes of the original Gabor jets compared to the compressed Gabor jets (independent of the image size). We use the mean squared error *MSE* to evaluate distortion of each Gabor jet:

$$MSE = \sqrt{\frac{1}{LDN} \sum_{j=1}^{LD} \sum_{k=1}^{N} \left(a_j^k - \hat{a}_j^k \right)^2} \quad (12)$$

where $\hat{a}_j^k$ is the *j*th compressed coefficient magnitude at landmark *k*.

4.1. Distortion Impact on Classification

As discussed earlier, standard rate-distortion performance measures are not sufficient to gauge the impact of the compression on classification. We thus devised the following experiment using a database of 800 images containing two frontal views of 400 different individuals. One half of the database was left uncompressed and treated as the central server database. The other 400 images were treated as client faces. In a second experiment, we explore feature compression for storage space reduction. In this case compressed features rather than facial images represent the database. This experiment can be regarded as the reverse of the distributed case.

The model graphs corresponding to the client faces were compressed using the embedded coding technique described above. The similarity was then computed between each compressed model graph and the model graphs of each face in the central database using Eq. 5. The recognition rate was obtained using:

$$P(Match \le X) = \frac{\text{No. of graph models matched within } X \text{ tries}}{\text{No. of images in the database}} \quad (13)$$

For evaluation purposes we compared the rate-distortion performance of the embedded quantizer to a standard scalar quantizer. We also compared the performance of the feature compression using embedded coding to the performance obtained by compressing the images using JPEG2000 prior to feature extraction.

5. Results

Figure 3 shows the rate distortion performance of the embedded and scalar quantizers (EQ and SQ, respectively) averaged over all images in the database.

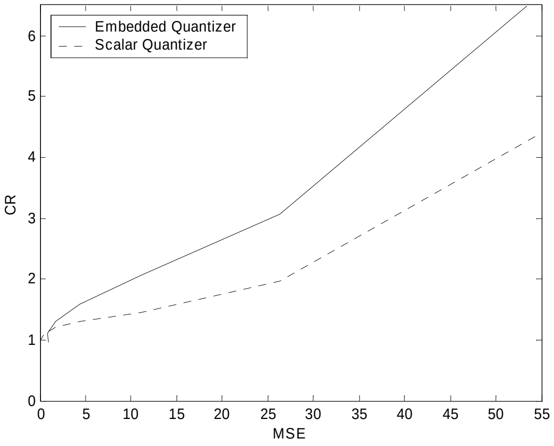


Figure 3. Rate-distortion performance.

Clearly, the embedded coder achieves better compression at equal distortion. This is encouraging in that it shows that the embedded coder is taking advantage of the Gabor wavelet structure. EQ achieved a maximum compression ratio of roughly 6 to 1 compared to 4 to 1 for SQ. Still, these results do not provide any indication of how the incurred distortion affects the face recognition performance. Furthermore, it is unclear what the compression ratio represents relative to the original images. The entropy calculated using Eq. (10) measures bits per landmark coefficient. In order to compare the rate performance of the embedded coder to image coding performance, the entropy *H* was scaled to bits per pixel (bpp):

$$H_{bpp} = H \frac{LDN}{I_1 I_2} \quad (14)$$

Where, *LD* is the number of wavelet coefficients, *N* is the number of landmarks, and I_1 and I_2 are the original image dimensions (128 x 128 pixels in this case). Figure 4 shows the rate distortion performance of EQ and SQ in terms of the bit-rate obtained using Eq. (14).

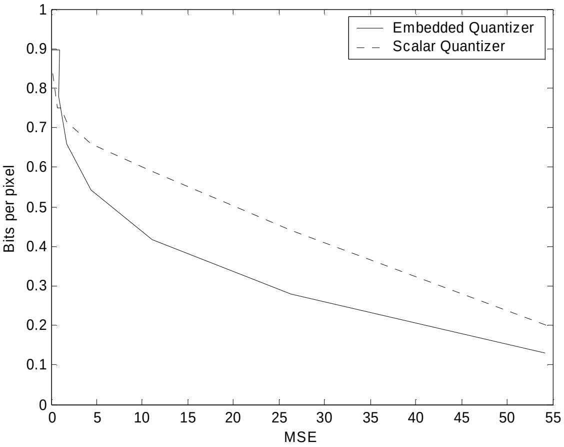


Figure 4. Scaled rate-distortion performance.

We have chosen to compare our proposed compression scheme with JPEG2000 (we used the version developed at EPFL, Switzerland). We evaluate the performance of both EQ and JPEG2000 in two scenarios. First, we consider the distributed case, where the features are extracted remotely, compressed and transmitted. Second, we consider the limited storage case, where features are compressed to reduce the storage burden. In this first case, we obtain recognition rates for compressed features compared to an uncompressed database. In the second case, we obtain recognition rates for uncompressed features compared to a compressed database.

Figure 5 shows the recognition rate of EQ and JPEG2000 vs. bit-rate for the distributed case. Figure 6 shows the recognition rate of EQ and JPEG2000 vs. bit-rate for the storage case.

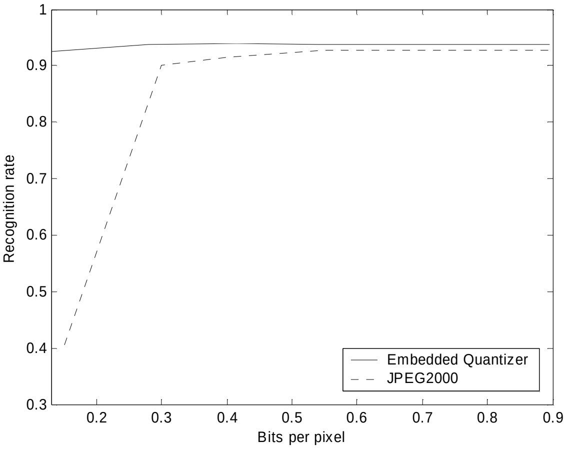


Figure 5. Recognition rates for the distributed case.

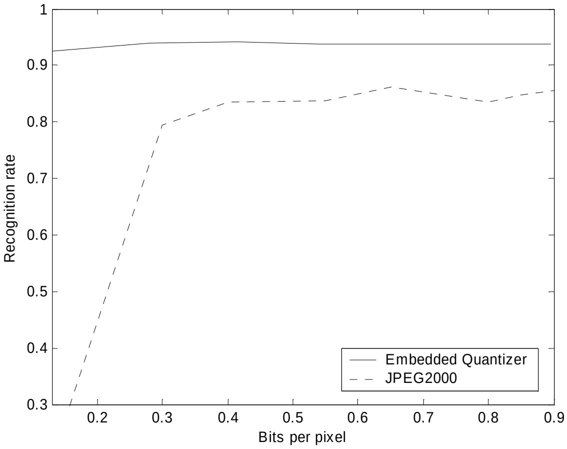


Figure 6. Recognition rates for the storage case.

As can be seen from Figures 5 and 6, the performance of EQ is much better than JPEG2000. In both the distributed and storage cases, EQ achieved a higher face recognition rate at an equal bit-rate. In fact, for the storage case, the recognition rates obtained using EQ are notably higher than those obtained using JPEG2000. This is noteworthy in that it is not uncommon to store a database of images using JPEG2000. However, our results indicate that it would be more advantageous to instead store the compressed features, as suggested here.

It should be noted that the recognition rates shown in Figure 5 and 6 are for when the face is matched exactly on the first try, i.e. for *X*=1 in Eq. (13). We also consider other criteria for success, such as matching a face within other values of *X*. These results are shown in Tables 1 through 4.

The face recognition rates were evaluated for the three lowest bit-rates in (bits per pixel) using Eq. (13). The recognition rates using EQ for the distributed case are shown in Table 1 compared to the recognition rates using only uncompressed features (with infinite precision Gabor jets). The bit-rates shown in Table 1 were averaged over all images in the database. The recognition rates obtained for model graphs extracted from JPEG2000 compressed images is shown in Table 2. The images were compressed at the target bit-rates shown in Table 2, which are very close to those obtained using EQ.

Table 1. Recognition rates using EQ (Distributed)

$P(Match \leq X)$	Uncompressed	0.13 bpp	0.27 bpp	0.41 bpp
X=1	93.8%	92.5%	93.8%	94.0%
X=2	95.0%	92.8%	94.3%	95.3%
X=5	95.5%	93.5%	95.0%	95.8%

Table 2. Recognition rates using JPEG2000 (Distributed)

$P(Match \le X)$	Uncompressed	0.15 bpp	0.30 bpp	0.40 bpp
X=1	93.8%	40.5%	90.0%	91.5%
X=2	95.0%	44.5%	90.5%	92.3%
X=5	95.5%	51.8%	91.5%	93.8%

Table 3. Recognition rates using EQ (Storage)

$P(Match \le X)$	Uncompressed	0.13 bpp	0.27 bpp	0.41 bpp
X=1	93.8%	92.5%	94.0%	94.3%
X=2	95.0%	93.3%	94.3%	95.0%
X=5	95.5%	94.8%	95.3%	95.5%

Table 4. Recognition rates using JPEG2000 (Storage)

$P(Match \le X)$	Uncompressed	0.15 bpp	0.30 bpp	0.40 bpp
X=1	93.8%	27.5%	79.5%	83.5%
X=2	95.0%	31.3%	80.0%	84.8%
X=5	95.5%	36.0%	81.8%	86.3%

As can be seen from Table 1, the embedded coding for feature compression impacts recognition rates minimally. Even at the lowest bit-rate (equivalent to a 6 to 1 compression ratio) there is an average decrease in classification performance of only 1%. At the higher bit-rates, the recognition rates are equivalent. The impact of the result is most notable when compared to the performance using JPEG2000 image compression on the images prior to feature extraction (Table 2). At the lowest bit-rate, the recognition rates are very poor. The classification rates increase significantly for higher bit-rates but are still below the embedded coder rates. The dramatic impact on recognition when compressing images at low bit-rates can be seen in Figure 7, which shows as example an uncompressed face image drawn from the database, together with the compressed images at the target bit-rates shown in Table 2. Clearly, the face in the image compressed at 0.15 bpp is unidentifiable.

Tables 3 and 4 show the results for the storage case. Again, in this case, the database is compressed. In the case of EQ, the compressed features are stored, whereas in the case of JPEG2000, the compressed images are stored and features are extracted from the compressed images. The difference in performance between EQ and JPEG2000 is more visible in Tables 3 and 4. There is virtually no difference between Tables 1 and 3. However, there is a further drop in JPEG2000 performance from Table 2 to Table 4.

The fact that the EQ compressed feature recognition rates on average drop by only 1% compared to the uncompressed features, suggests that the structure of the features is being adequately preserved and large gains in compression can be made without a significant impact in classification performance. Furthermore, it is clear that our proposed feature compression scheme can be used with equal success for distributed face recognition applications and also for storage savings, whereas this is not the case for JPEG2000.

Finally, in comparing the results of Tables 1 through 4, the conclusion can be drawn that compressing feature vectors as opposed to images is preferable in distributed classification applications, as well as for reduced storage impact.

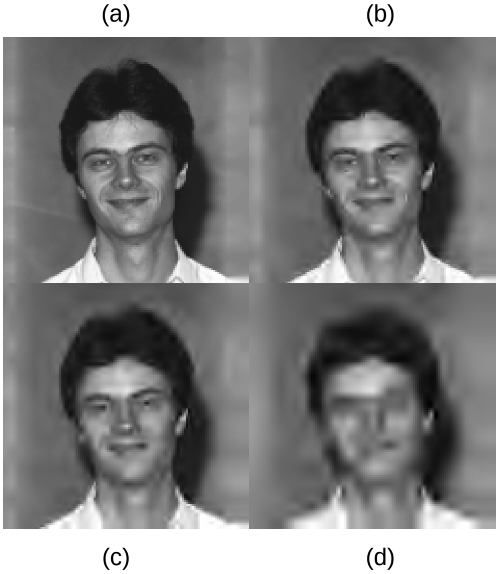


Figure 7. Original (a), 0.40 (b), 0.30 (c), and 0.15 bpp (d).

6. Conclusions and Future Work

We addressed compression for distributed face recognition by investigating the impact of feature compression on overall face recognition rates. Given that the Bochum/USC face recognition system employs Gabor wavelet features, we propose using a modified embedded coding scheme. Our evaluation showed that the embedded coder achieves a bit-rate as low as 0.13 bpp with a minimal impact on recognition rates (a 1% decrease on average). Furthermore, our evaluation showed that the classification performance is significantly better when compressing feature vectors compared to the classification performance obtained from features extracted from compressed images. In addition to the distributed face recognition case, we also investigated the representation of a database of faces using compressed feature vectors. Our results showed that significantly higher recognition rates could be achieved using our proposed compression scheme vs. the state of the art JPEG2000 compression standard.

Given these promising results, we believe it will be worthwhile to study variable coding rates for the facial landmarks. The variable rates could be determined using existing knowledge obtained from previous studies of the Bochum/USC face recognition system with parametric linear subspaces [6].

7. References

[1] N. Srinivasamurthy, A. Ortega and S. Narayanan, "Towards Optimal Encoding for Classification with Applications to Distributed Speech Recognition," *Eurospeech 2003*, Geneva, Switzerland September 2003.

[2] H. Xie and A. Ortega, "Entropy- and Complexity-constrained Classified Quantizer Design for Distributed Image Classification," *IEEE International workshop on Multimedia Signal Processing*, Virgin Islands, December 2002.

[3] S. S. Pradhan, J. Kusuma, and K. Ramchandran, "Distributed Compression in a Dense Microsensor Network," *IEEE Signal Processing Magazine*, March 2002.

[4] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The Bochum/USC Face Regognition System and How it Fared in the FERET Phase III Test," *Face Recognition: From Theory to Applications*, H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelman Soulie, T.S. Huang (Eds.). Springer-Verlag, pp.186-205, 1998.

[5] J. M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients," *IEEE Trans. Signal Processing*, vol. 41, December 1993.

[6] L Wiskott, J.-M. Fellous, N. Krueger, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 19, 1997.

[7] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 22, 2000.

[8] K. Okada and C. von der Malsburg, "Pose-Invariant Face Recognition with Parametric Linear Subspaces," *Fifth International Conference on Automatic Face and Gesture Recognition*, Washington DC, May 2002.

[9] S. G. Mallat, "A Theory of Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, July 1989.

[10] D. S. Taubman and M. W. Marcellin, "JPEG2000: Image Compression Fundamentals, Standards and Practice," Kluwer Academic Publishers, 2002.

TOWARDS STILL TO VIDEO BASED FACE RECOGNITION

S. Palanivel & B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engg.
Indian Institute of Technology Madras
Chennai-600 036, India.

Chunyan Xie & B.V.K. Vijaya Kumar

Department of Electrical and Computer Engg.
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A.

ABSTRACT

The objective of this paper is to address the issues involved in deriving the evidence from a video sequence of images of a face when they are matched with the static image of the face available as a reference. This problem arises in the context of face identification and verification. The test data consists of a sequence of face images of a naturally moving person, captured from one or more cameras. An automatic face recognition system involves face/head tracking, normalizing the face region, representation and matching of test and reference images to derive evidence, and combining the evidence from multiple frames of the face images. The head contour points of the moving person are extracted using the motion information. The face region is extracted and normalized to account for scaling and orientation to derive the normalized face image. The normalized face image or some selected portion of the face image is matched with the reference image to derive partial evidence. The partial evidence obtained from each frame in the video sequence is combined to decide the identity of the person.

1. INTRODUCTION

Automatic face identification or verification by machine appears to be difficult, while it is done effortlessly by a human being. The main reason for this difficulty is that it is difficult to articulate the mechanism humans use. For machine recognition of faces, simplified assumptions are made in the feature extraction and matching, and the face images are captured under restricted and severely constrained environment. For example, most face recognition studies assume the availability of the cropped up face image so that difficulties due to variation in scale and orientation are minimized [1]. Likewise, except for a few carefully designed databases the face image data is collected mostly for frontal pose, so that effects of pose variation can be ignored. Effects of illumination, shadows and other lighting conditions are also reduced by collecting the data under controlled environment.

This paper address the issues involved in deriving the evidence from a video sequence of images of a face when they are matched with the static image of the face available as a reference. The test data consists of face images of a naturally moving person, captured from one or more cameras. It is possible that none of the captured images may contain a face image suitable for matching directly with the reference image. However, there may be some images in the video test sequence which may give a good match with the reference image, provided proper representation and matching methods are available. Combining the evidence from these frames may lead to a better decision for recognizing the person in the video. Since the illumination conditions during test are usually different from those with the reference collection, matching of the images may not result in high confidence values even for the authentic case. Moreover, the changes due to expression, pose and other variations make the problem of matching a test image with the reference challenging. It may be necessary to derive evidence by matching selected parts of the reference face image with the corresponding parts in the test image, and then combine the partial evidence to derive the evidence at the frame level matching.

Matching two face images require representation of the image. For cropped up images, eigenvectors are derived, and the first few components of an image projected onto these eigenvectors are used to represent the image for matching [2]. Several variations of the eigenvector approach are available in the literature [3],[4]. Other matching methods based on elastic graph representation and statistical distribution of the facial features have also been investigated in the literature for face recognition. [5]-[7].

The key elements in the face recognition problem are representation and matching of face images. We discuss the issues involved in the context of recognizing or verifying a given still face image in a sequence of images of the face collected by a camera. Characteristics of the reference and test images, and the issues in developing a face recognition system are discussed in the next section. Subsequent sections discuss each of these issues in more detail. In Section

3, we consider the issue of tracking and normalization of the face region from a video. The effects of matching and deriving the partial evidences are discussed in Section 4. The partial evidence obtained from each frame in the video sequence can be combined using an Autoassociative neural network (AANN) model, and it is discussed in Section 5.

2. CHARACTERISTICS OF REFERENCE AND TEST DATA

The reference data consists of one digital image of the face for each person. The reference face image of a person is generally of high quality with cropped up (manually if needed) region of the face. The test data consist of video sequences of a single person walking normally in a specified zone where it may be possible to collect video sequences with more than one video camera, covering different views/angles, if necessary. The test subject may not be cooperative, in terms of giving the specific views to the camera, and also the subject may have different facial appearance including make up. The illumination also may be varying. Some of the frames in the video may not have any portion of the frontal face, and may not have even the face in the field of the camera. In such a scenario, the objective is to determine whether the person in view is same as the person in one of the reference images (identification), or is same as the person in a specific reference image whose identity we want to verify from the video (verification). It is obvious that only a few of the frames in the video sequence may have a view approximately corresponding to the frontal view of the face of the person. It is those frames that are likely to provide high confidence value with the reference face image, provided suitable representation and matching are available.

The following stages are involved in developing an algorithm for matching the video test sequence with a reference image:

(a) Face tracking and normalization,

(b) Matching a test and reference face image and deriving the evidence,

(c) Combining the evidence.

3. FACE TRACKING AND NORMALIZATION

Computer vision based applications such as face recognition requires automatic detection and tracking of human head or face in an image sequence. However, many applications in the literature assume that the faces in the image sequence have been localized.

Methods has been proposed in the literature for head tracking based on intensity gradients and color histograms [8], statistical model of color and shape [9], 3D modeling [10], temporal information [11], Gaussian Mixture Model [12], Kalman filter [13]. The method proposed in this paper uses the motion information to extract the head contour points.

3.1. Extracting head contour points

The head contour points are extracted from the gray level interframe difference image. The RGB image is converted to gray level image (I), and the interframe difference image (D) is obtained by

$$\begin{aligned} D(i,j,k) &= |I(i,j,k) - I(i,j,k-1)| \\ & \quad 0 \leq i < w, 0 \leq j < h, k \geq 1 \end{aligned} \tag{1}$$

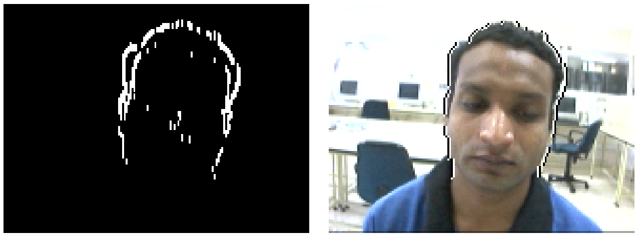
where k is the frame number in the video, w and h are the width and height of the image, respectively.

The thresholded difference image T is obtained by

$$T(i,j,k) = \begin{cases} 1, & if\ D(i,j,k) > \lambda \\ 0, & otherwise \end{cases} \tag{2}$$

where λ is the threshold, which is the smallest integer such that $T(i,j,k) = 0$, for all i and j, whenever there is no moving region in the camera view.

The thresholded difference image is scanned from top to bottom to find out an approximate top pixel (c_x, c_y) of the moving region. The head contour points are extracted by scanning the thresholded difference image from the pixel (c_x, c_y). This process is repeated for every two consecutive frames in order to track the face region in the video. Fig.1(a) shows the thresholded difference image as given by the Eq.2 and Fig.1(b) shows the extracted head contour points.



(a) Difference image (b) Contour points

Fig. 1. Head contour points.

3.2. Fitting an ellipse

The method proposed in [14] is used to fit an ellipse for the extracted head contour points. In this method a generic conic is represented as the zero set of an implicit second order polynomial as given in Eq.3. If F is a function on an open set U, then the zero set of F is the set $Z = \{z \in U : F(z) = 0\}$.

$$F(\mathbf{a},\mathbf{x}) = \mathbf{ax} = ax^2 + bxy + cy^2 + dx + ey + f \quad (3)$$

where $\mathbf{a}$ = [*a b c d e f*] and $\mathbf{x}$ =[x^2 xy y^2 x y 1]^T. $F(\mathbf{a}, \mathbf{x_i}) = d_i$ is called the ”algebraic distance” of a point $\mathbf{x_i}$ to the conic $F(\mathbf{a}, \mathbf{x}) = 0$.

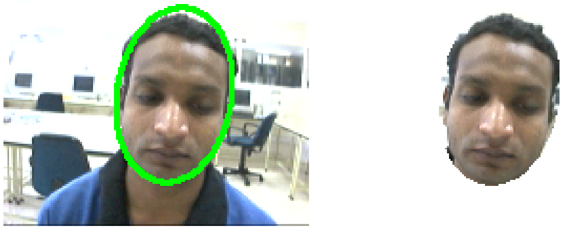
One way of fitting a conic is to minimize the algebraic distance over the set of N data points in the least squares sense as given in Eq.4.

$$\hat{\mathbf{a}} = arg \min_{\mathbf{a}} \left\{ \sum_{i=1}^{N} F(\mathbf{a}, \mathbf{x_i})^2 \right\} \quad (4)$$

The method proposed in [14] minimize the algebraic distance given in Eq.4 subject to the constraint $b^2 - 4ac < 0$. The method gives the center, width, height and angle of the ellipse for the given N contour points of the head. The width (e_w) and height (e_h) of the ellipse for the head region satisfies the constraint given in Eq.5.

$$1.4 * e_w \leq e_h \leq 1.9 * e_w \quad (5)$$

The Bresenham’s ellipse generation algorithm is used to generate the ellipse using the estimated center, width, height and angle values [15]. The generated ellipse and the face region are as shown in Fig.2.



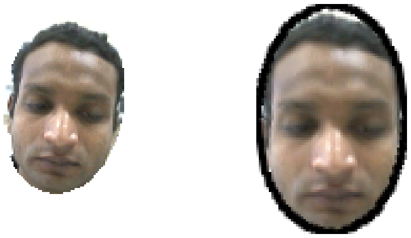
(a) Generated ellipse (b) Face region

Fig. 2. Fitting an ellipse.

The proposed method requires initial head movement. If there is no motion in the successive frames, then the previous ellipse coordinates are retained. Experimental results show that this method is invariant to scaling, illumination and facial expressions. It is also invariant to tilt, yaw and pose of the face or head to some extent.

3.3. Face normalization

The elliptic face region obtained from the video is normalized to account for scaling and orientation to derive the normalized face image as shown in Fig.3. The width and height of the ellipse is used to normalize the elliptic face region to a fixed size, and the angle is used to normalize the orientation. Figs.3(a) and 3(b) show the unnormalized and normalized face images, respectively. Fig.4 shows the result of face tracking for two subjects and the corresponding normalized face images are shown in Fig.5.



(a) Unnormalized face (b) Normalized face

Fig. 3. Face normalization.

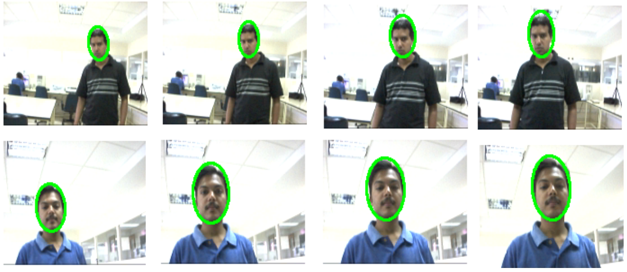


Fig. 4. Face tracking

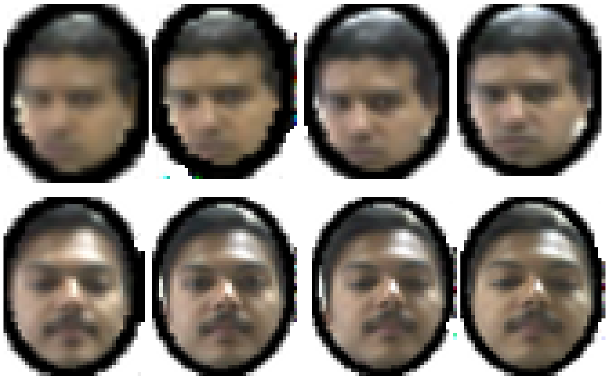


Fig. 5. Normalized face images

An effective method of comparing a test and reference image is by using correlation filters [16]. An important metric in the use of correlation filters is the peak-to-sidelobe ratio (PSR) which quantifies the sharpness of the correlation peak. For well-designed correlation filters, PSR should be large for authentics and small for impostors. Using PSR of the correlation output one can derive the evidence for a sequence of frames of test images and a reference face image. The PSR is defined as

$$PSR = (p - \mu)/\sigma \tag{6}$$

where p is the peak value in the correlation output, μ and σ are the mean and the standard deviation in a side-lobe region, excluding a 5×5 mask centered at the peak. The size of the side-lobe region is typically 20×20 for a 64×64 face image. These window sizes are empirically derived and other choices may be better for other cases. The PSR plots for a genuine test sequence (subject 1) and a few impostors test sequences are shown in Fig.6. The thick line shows the PSR values for the genuine test sequence.

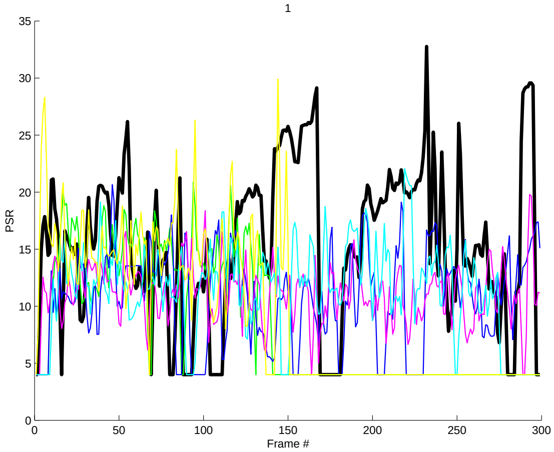


Fig. 6. PSR plot for genuine (subject 1) and impostor test sequence

But despite the exploitation of the behavior of the correlation filtering, it may not be possible most of the time to obtain good match between the test and reference face images, due to pose, tilt, illumination differences and partial visibility of the face in a video frame. Fig.7 shows the PSR plots for a genuine test sequence (subject 2) and a few impostors test sequences. In this case most of the time the genuine PSR values are lower when compared to the plots for the impostors. In such cases it is worth exploiting the evidence obtained from selected portions of the reference image as discussed in the next section.

4. MATCHING A TEST AND REFERENCE FACE IMAGE

If a small portion of the face image is used for correlation matching, it is likely that the random variations in the pixel values may result in a poor match even for the authentic case. Therefore it is better to represent only the significant part of the face image, by reducing variations due to noise. For this we propose a representation of the face image using 1-D eigenvectors. These vectors are derived from the reference image using the columns of the image pixels as 1-D vectors. Fig.8 shows the image obtained using the first 15 eigenvectors derived from the corresponding reference image. The reconstructed image reduces the variations due to noise, although some blurring is also present. The selected portions of the reconstructed reference image such as eyes, nose-mouth and mouth parts as shown in Fig.9 can be matched with the test image to derive partial evidence.The test image also is represented in terms of the first 15 eigenvectors of the reference image. It may be possible to select the portions of the reference image which need not correspond to any specific features like eyes, mouth, nose etc. For example, one can use some columns or rows of the image pixels for deriving the partial evidence.

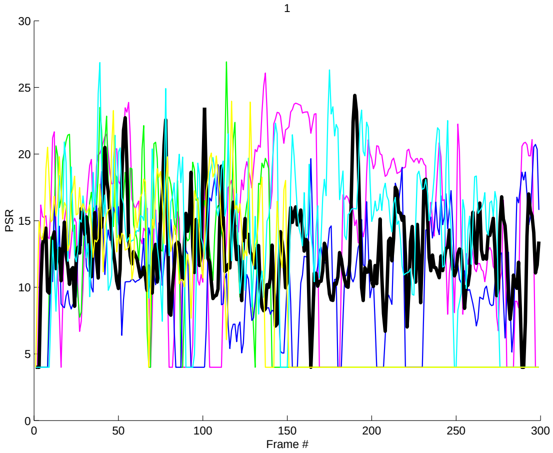


Fig. 7. PSR plot for genuine (subject 2) and impostors test sequence

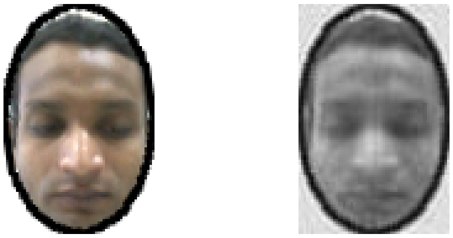

(a) Face image
(b) Reconstructed face image

Fig. 8. 1-D eigenvector representation of face image.

Figs.10 and 11 show the evidence obtained from eyes, nose-mouth portion of the face image for the test subject 1 and 2, respectively. The genuine test sequence PSR values are represented by square and the impostors PSR values are represented by plus. Figs.12 and 13 show the evidence when the three features, eyes, nose-mouth and mouth (as shown in Fig.9) are used.

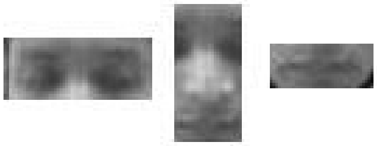


Fig. 9. Eyes, nose-mouth and mouth portion of the reconstructed reference image

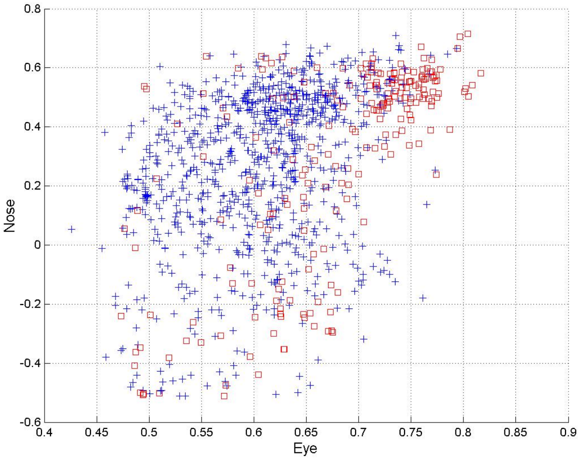


Fig. 10. Partial evidence for test subject 1

5. COMBINING THE EVIDENCE

The evidences obtained from the selected portions from each frame in the video test sequence can be combined using an autoassociative neural network (AANN) model. AANN model is a feedforward neural network performing an identity mapping of the input space. It can be used to capture the distribution of the input data [17],[18]. The distribution of the impostor evidence for each reference subject is captured using a five layer AANN model. Fig.14 shows the structure of the AANN model used in our study. It can be denoted as 3L 6N 2N 6N 3L, where L denotes a linear unit, and N denotes a nonlinear unit. The integer value indicates the number of units used in that layer. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the first or fifth. The

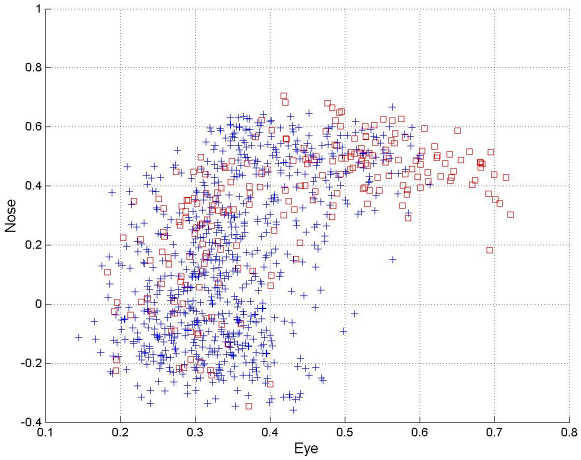


Fig. 11. Partial evidence for test subject 2

activation functions at the second, third and fourth layer are nonlinear. The nonlinear units use $tanh(s)$ as the the activation function, where s is the activation value of the unit. The standard backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

The evidences are derived from each frame in the video test sequence with respect to a reference subject. These evidences are given as input to the model corresponding to the reference subject. The output of the model is compared with the input to compute the normalized squared error. This squared error gives an indication of the confidence with which the input frame belongs to the impostor class. The smaller the error, the higher the confidence with which we may label the input to belong to the impostor class. Therefore, it seems logical to assume that larger error gives an indication of the confidence with which the input can be assigned to the authentic class. Although it is desirable to derive a suitable confidence measure from these error values, in this paper we have used the error value itself as the confidence value for the authentic class. The accumulated error is calculated from the error obtained for each frame in the video sequence. This process is repeated for all the reference subjects. The largest accumulated error is used to decide the identity of the test subject. The result of the accumulated error for the test subject 1 with respect to its model is given in Fig.15. The thick line corresponds to the accumulated error with respect to the reference subject 1. Fig. 16 shows the corresponding plot for the subject 2. As can be seen from the plot, as more frames are used to accumulate the error for the model of the subject, it will exceed the error from all other models for the genuine case. Thus the evidence collected from selected portions can be

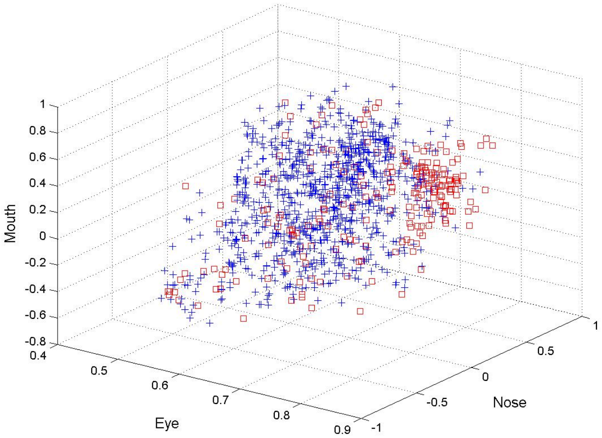


Fig. 12. Partial evidence for test subject 1

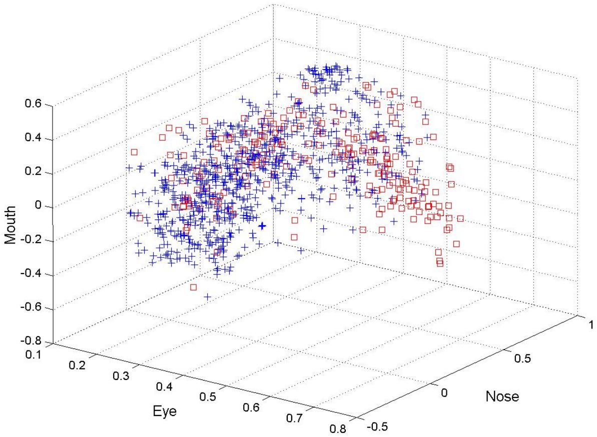


Fig. 13. Partial evidence for test subject 2

combined and accumulated to derive a better decision from the video test sequence. Note that the evidence for the authentic is significantly better in these plots compared to the evidence obtained by direct correlation matching shown in Figs.6 and 7. In particular, note that the evidence for subject 2 has significantly improved as compared to the evidence in Fig.7. It may be possible to enhance the evidence further by selecting the subset of similar faces from the plots and using other clues, such as other parts of the face image and the knowledge of the video sequence.

6. CONCLUSION

This paper proposed a method for extracting the face region using the motion information. The extracted face image is normalized with respect to scale and orientation. The selected portion of the normalized face image is matched with the reference image to derive the partial evidence. The partial evidences obtained from each frame in the video sequence are combined using an autoassociative neural network model. The proposed method can be used effectively for matching a video test sequence with a still reference image.

7. REFERENCES

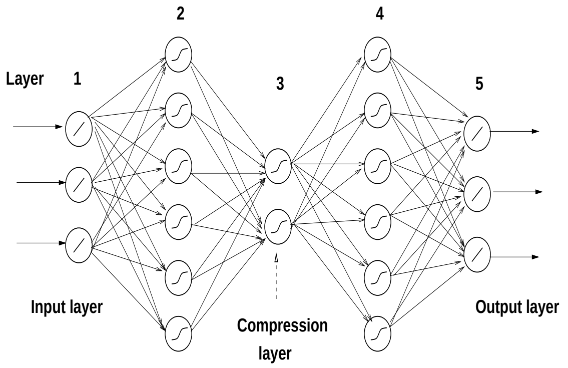


Fig. 14. AANN model used for capturing the distribution of input data.

[1] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, *Face Recognition: A Literature Survey*, UMD CAFR, Technical Report, CAR-TR-948, October 2000.

[2] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive NeuroScience*, vol. 3, pp. 71–86, 1991.

[3] Alper Yilmaz and M.Gokmen, “Eigenhill vs. eigenface and eigenedge,” *Pattern Recognition*, vol. 34, pp. 181–184, 2001.

[4] S. Ramesh, S. Palanivel, Sukhendu Das, and B. Yegnanarayana, “Eigenedginess vs. eigenhill, eigenface and eigenedge,” in *European Signal Processing Conference*, Toulose, France, September 3-6 2002, pp. 559–562.

[5] Constantine L. Kotropoulos, Anastasios Tefas, and Ioannis Pitas, “Frontal face authentication using discriminating grids with morphological feature vectors,” *IEEE Transactions on MultiMedia*, vol. 2, no. 1, pp. 14–26, March 2000.

[6] Constantine Kotropoulos and Ioannis Pitas, “Using support vector machines to enhance the performance

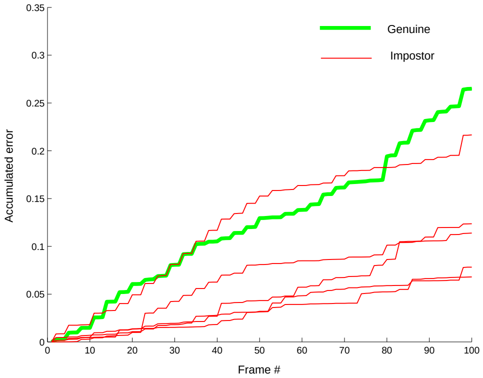


Fig. 15. Accumulated error for test subject 1

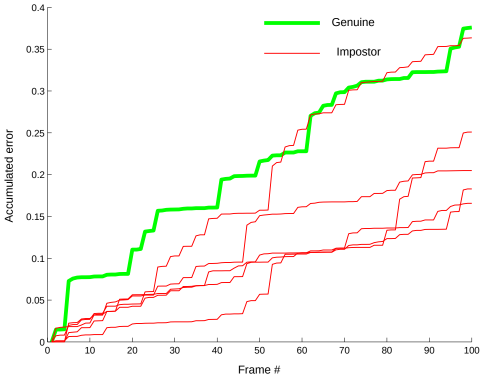


Fig. 16. Accumulated error for test subject 2

of elastic graph matching for frontal face authentication," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 7, pp. 735–746, July 2001.

[7] S. Palanivel, B. S. Venkatesh, and B. Yegnanarayana, "Real time face authentication system using autoassociative neural network models," in *IEEE International conference on Multimedia and Expo*, Baltimore, July 2003, pp. 257–260.

[8] Stan Birchfield, "Elliptical head tracking using intensity gradients and color histograms," in *Poceedings of the IEEE conference on computer vision and pattern recognition*, Santa Barbara, California, June 1998, pp. 232–237.

[9] Christopher Richard Wren, Ali Azarbayejani, Trever Darrell, and Alex Paul Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, July 1997.

[10] Ioannis Kakadiaris and Dimitris metaxas, "Model-based estimation of 3d human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1453–1459, December 2000.

[11] Yann Ricquebourg and patrick bouthemy, "Real-time tracking of moving persons by exploiting spatio-temporal image slices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 797–808, August 2000.

[12] Tsuhan Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, pp. 9–21, January 2001.

[13] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Signal Processing Magazine*, pp. 22–31, January 2001.

[14] Maurizio Pilu, Andrew W. Fitzgibbon, and Robert B. Fisher, "Ellipse-specific direct least-square fitting," in *IEEE International conference on Image processing*, Lausanne, September 1996.

[15] W.M. Newmann and R.F.Sproull, *Principles of interactive Computer Graphics*, McGraw Hill, New York, 1979.

[16] B.V.K. Vijaya Kumar, Marios Savvides, Krithika Venkataramani, and Chunyan Xie, "Spatial frequency domain image processing for biometric recognition," in *IEEE International conference on Image Processing*, New York, September 2002, pp. 53–56.

[17] B. Yegnanarayana and S.P. Kishore, "AANN: an alternative to GMM for pattern recognition," *Neural Networks*, vol. 15, pp. 459–469, January 2002.

[18] B. Yegnanarayana, Suryakanth V. Gangashetty, and S. Palanivel, "Autoassociative neural network models for pattern recognition tasks in speech and image," in *Soft Computing Approach to Pattern Recognition and Image Processing*, World Scientific publishing Co. Pte. Ltd, Singapore, December 2002, pp. 283–305.

Visible-light and Infrared Face Recognition

Xin Chen Patrick J. Flynn Kevin W. Bowyer
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
{xchen2, flynn, kwb}@nd.edu

Abstract

This study examines issues involved in the comparison and combination of face recognition using visible and infra-red images. This is the only study that we know of to focus on experiments involving time lapse between gallery and probe image acquisitions. Most practical applications of face recognition would seem to involve time-lapse scenarios. We find that in a time lapse scenario, (1) PCA-based recognition using visible images may outperform PCA-based recognition using infra-red images, (2) the combination of PCA-based recognition using visible and infra-red imagery substantially outperforms either one individually, and (3) the combination of PCA-based recognition using visible and infra-red also outperforms a current commercial state-of-the-art algorithm operating on visible images. For example, in one particular experiment, PCA on visible images gave 75% rank-one recognition, PCA on IR gave 74%, FaceIt on visible gave 86%, and combined PCA IR and visible gave 91%.

1 Introduction

Face recognition in the thermal domain has received relatively little attention in the literature in comparison with recognition in visible imagery. This is mainly because of the lack of widely available IR image databases. Previous work in this area shows that well-known face recognition techniques, for example PCA, can be successfully applied to IR images, where they perform as well on IR as on visible imagery [1] or even better on IR than on visible imagery [2] [3]. However, in all of these studies [1] [2] [3], the gallery and probe images of a subject were acquired in the same session, on the same day. In our current study, we also examine performance when there is substantial time elapsed between gallery and probe image acquisition.

Socolinsky and Selinger [2] [3] used 91 subjects, and the gallery and probe images were acquired within a very short period of time. We will refer to such experiments as *same session recognition*. Experiments in which the probe and gallery images are acquired on different days or weeks will be called *time-lapse recognition*. Socolinsky and Selinger used a sensor capable of imaging both modalities (visible and IR) simultaneously through a common aperture. This enabled them to register the face with reliable visible images instead of IR images. They emphasized the IR sensor calibration and their training set is the same as the gallery set. In their experiments, several face recognition algorithms were tested and the performance using IR appears to be superior to that using visible imagery.

Wilder *et al.* [1] used 101 subjects and the images were acquired without time lapse. They controlled only for expression change. Several recognition algorithms were tested and they concluded that the performance is not significantly better for one modality than for another.

Additional work on IR face recognition has been done by [4] and [5][6]. In [4], an image data set acquired by Socolinsky *et al.* was used to study multi-modal IR and visible face recognition using the Identix FaceIt algorithm [7]. In [5][6], IR face recognition was explored with a smaller dataset, but combined IR and visible images for face recognition was not addressed.

This study examines more varied conditions and uses a relatively larger database, in both the number of images and the number of subjects, compared with the databases used by Wilder *et al.* and Socolinsky *et al.* [1] [2] [3]. We consider the performance of the PCA algorithm in IR, including the impact of illumination change, facial expression change and the short term (minutes) and longer term (weeks) change in face appearance. This current work is an extension of previous work [8] to more carefully consider the relative effects of time lapse between gallery and probe images on the performance of infrared versus visible imagery, and also to investigate the accuracy of eye center location as a possible cause for the inferior performance of infrared relative to visible-light images in a time-lapse scenario.

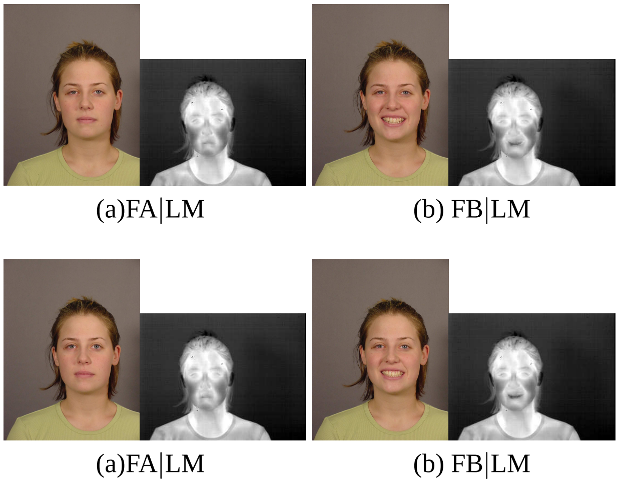
2 Data Collection

Most of the data used to obtain the results in this paper was acquired at University of Notre Dame during 2002, where IR images from 240 distinct subjects were acquired. Each image acquisition session consists of four views with different lighting and facial expressions. Image acquisitions were held weekly for each subject and most subjects participated multiple times. All subjects completed an IRB-approved consent form for each acquisition session. IR images were acquired with a Merlin [1] Uncooled long-wavelength IR camera, which provides a real-time, 60Hz, 12 bit digital data stream, has a resolution of 320×240 pixels and is sensitive in the 7.0-14.0 micron range. Visible-light images were taken by a Canon Powershot G2 digital camera with a resolution of 1200×1600 and 8 bit output. Three Smith-Victor A120 lights with Sylvania Photo-ECA bulbs provided studio lighting. The lights were located approximately eight feet in front of the subject. One was approximately four feet to the left, one was centrally located and one was located four feet to the right. All three lights were trained on the subject's face. The side lights and central light are about 6 feet and 7 feet high, respectively. One lighting configuration had the central light turned off and the others on. This will be referred to as "FERET style lighting" or "LF". The other configuration has all three lights on; this will be called "mugshot lighting" or "LM". For each subject and illumination condition, two images were taken: one is with neutral expression, which will be called "FA", and the other image is with a smiling expression, which will be called "FB". For all of these images the subject stood in front of a standard gray background. Since glass and plastic lenses are opaque in IR, we asked all subjects to remove eyeglasses during acquisition. According to the lighting and expression, there are four categories: (a) FA expression under LM lighting (FA|LM), (b) FB expression under LM lighting (FB|LM), (c) FA expression under LF lighting (FA|LF) and (d) FB expression under LF lighting (FB|LF). Figure 1 shows one subject in one session under these four conditions.

To create a larger training set for our experiments, we also used 81 IR and visible-light images of 81 distinct subjects, acquired by Equinox Corporation [9].

3 Preprocessing

We located faces manually by clicking on the centers of each eye. The features on a human face are much more vague in IR than in visible imagery and thus the registration in the following normalization step might not be as reliable in IR as in the visible images. Notice that Socolinsky and Selinger [2] [3] used a sensor capable of capturing simultaneous registered visible and IR, which is of particular significance for their comparison of visible and IR. The fact that they get eye location from visible imagery and use it in IR may make their IR performance better than if they used IR alone for eye location.


(a)FA|LM

(b) FB|LM

(a)FA|LM

(b) FB|LM

Figure 1: Face images in visible and IR under different lighting and facial expression conditions.

A PCA subspace is derived separately for visible and IR images of the same 240 individuals. These individuals are not in the gallery or probe sets. We followed the convention in the CSU software [10] and used 130 x 150 resolution versions of the original visible and IR images in creating the face space. Recognition is performed by projecting a probe image into the face space and finding the nearest gallery image. The "MahCosine" metric is used to compute the distance between points in the face space [10].

4 Same-session Recognition

We used 82 distinct subjects and four images for each subject acquired within 1 minute with different illumination and facial expressions. For each valid pair of gallery and probe sets, we computed the rank 1 correct match percentage and the rank at which all the probes were correctly matched. They are reported in Table 1. Each entry in the leftmost column corresponds to a gallery set, and each entry in the top row corresponds to a probe set. The subspace for Table 1 was derived by using 240 images of 240 distinct subjects.

Table 1 shows that there is no consistent difference between the performance of visible and IR. IR is better in six instances, visible is better in four instances and they are the same in two instances. The overall performance for same session recognition is high for both IR and visible, and so it is possible that some "ceiling effect" could make it difficult to observe any true difference that might exist.

5 Time-lapse Recognition

Time-lapse recognition experiments use the images acquired in ten acquisition sessions of Spring 2002. In the ten acquisition sessions, there were 64, 68, 64, 57, 49, 56,

[1] Manufacturer names are given only to specify the experimental details more precisely, and not to imply any endorsement of a particular manufacturer's equipment.

Table 1: The percentage of correctly matched probes at rank 1 and the smallest rank at which all probes are correctly matched for same session recognition in Visible(bottom) and IR(top)

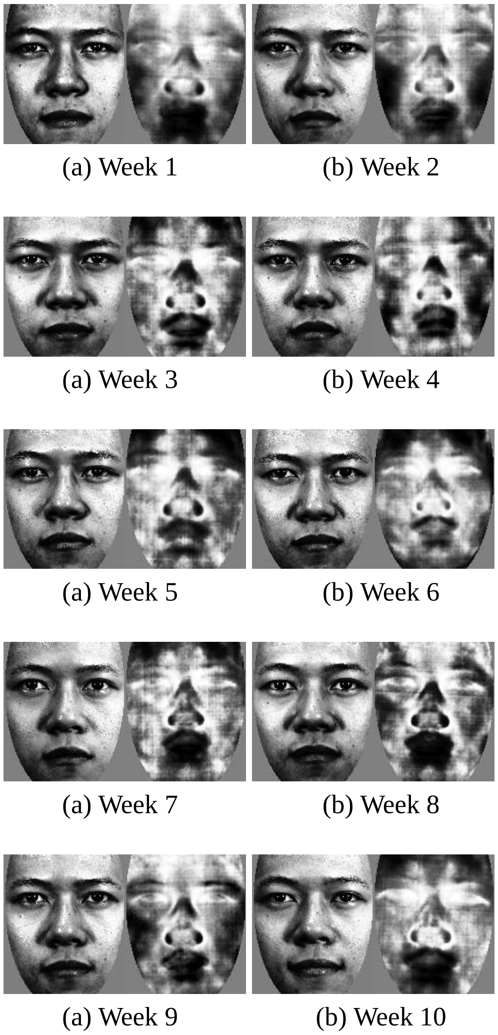
	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		0.98 (2) 0.98 (10)	0.99 (3) 0.98 (10)	0.99 (2) 0.94 (4)
FA|LM	0.99 (2) 0.95 (6)		0.94 (28) 1.00 (1)	0.95 (19) 1.00 (1)
FB|LF	0.96 (4) 0.95 (6)	0.95 (39) 1.00 (1)		1.00 (1) 1.00 (1)
FB|LM	0.98 (2) 0.89 (17)	0.96 (19) 0.98 (3)	1.00 (1) 0.98 (3)	

54, 54, 60, and 44 subjects. Figure 2 shows the visible and IR images of one subject across 10 different weeks, which suggests that there may be more apparent variability, on average, in the IR images of a person than in the visible images. In particular, the bridge and sides of the nose appear somewhat different in different IR images. [11] confirmed that there is variability in IR images due to startling, gum-chewing and walking exercise, etc.

The scenario for this recognition is a typical enroll-once identification setup. There are 16 experiments based on the exhaustive combinations of gallery and probe sets given the images of the first session under a specific lighting and expression condition as the gallery and the images of all the later sessions under a specific lighting and expression condition as the probe. That is, each gallery set has 64 images from session 1; each probe set has 431 images from sessions 2-10. The rank-1 correct match percentages are given in Table 2. For each subject in one experiment, there is one enrolled gallery image and up to nine probe images, each acquired in a distinct later session. The same face space is used as in the ”same-session” experiments.

Table 2: Rank 1 correct match percentage for time-lapse recognition in visible (bottom) and IR (top). Row indicates gallery and column indicates probe.

	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.83 (41) 0.91 (39)	0.84 (27) 0.93 (54)	0.77 (48) 0.73 (56)	0.75 (43) 0.71(56)
FA|LF	0.81 (38) 0.92 (31)	0.82 (46) 0.92 (28)	0.74 (49) 0.75 (32)	0.73 (43) 0.73 (44)
FB|LM	0.77 (45) 0.77 (33)	0.80 (49) 0.81 (44)	0.79 (39) 0.86 (48)	0.78 (51) 0.85 (47)
FB|LF	0.73 (58) 0.75 (41)	0.76 (58) 0.79 (40)	0.77 (36) 0.90 (27)	0.76 (41) 0.90 (47)



(a) Week 1 (b) Week 2

(a) Week 3 (b) Week 4

(a) Week 5 (b) Week 6

(a) Week 7 (b) Week 8

(a) Week 9 (b) Week 10

Figure 2: Normalized FA|LM face images of one subject in visible and IR across 10 weeks.

For IR, Table 2 illustrates a striking difference in performance in contrast to same-session recognition results shown in Table 1: the rank 1 correct match rate drops by 15% to 20%. The most obvious reason is that the elapsed time caused significant changes among thermal patterns of the same subject. In addition, it is possible that unreliable registration of the eye centers could have degraded the performance. Table 2 also shows that the performance degrades for visible imagery compared with that in same-session recognition. Visible imagery outperforms IR in 12 of the 16 cases, with IR and visible the same in another two.

For one time-lapse recognition with FA|LF images in the first session as the gallery set and FA|LF images in the second to the tenth sessions as the probe set, we illustrate the match and non-match distance distributions in Figure 3 and Figure 4. The score (distance) ranges from -1.0 to 1.0 since we use the ”MahCosine” distance metric in CSU software. The match score histogram is the distribution of distances between the probe images and their correct gallery matches. The non-match score histogram is the distribution of distances between the probe images and all their false gallery matches. Essentially, the match score distribution represents the within-class difference, while the non-match score distribution represents the between-class difference. Hence, for an ideal face recognition, the match scores should be as small as possible and the non-match scores should be much larger than the match scores and they shouldn't overlap. In this experiment, there is significant overlapping for both IR and visible-light, which accounts for the incorrect matches. The match score distribution for visible is more at the smaller distance area than that for IR, i.e., the within-class difference for visible is smaller than that for IR. The non-match score distributions for these two modalities are about the same, i.e., the between class differences are similar. Thus, visible-light imagery performs better than IR.

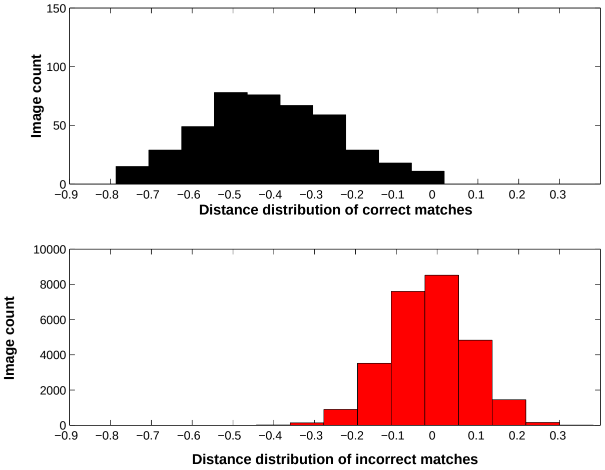


Figure 3: Match and non-match score distributions for one time-lapse recognition in IR

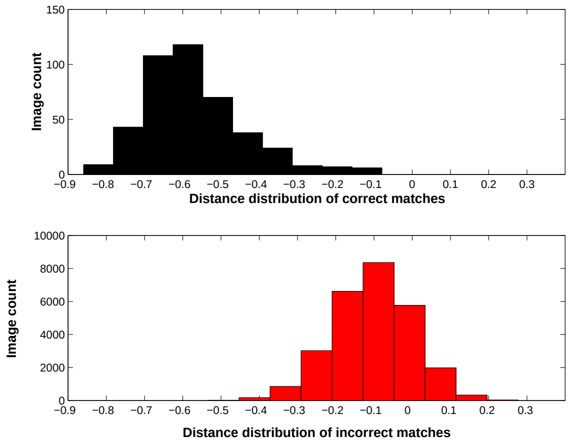


Figure 4: Match and non-match score distributions for one time-lapse recognition in visible-light

6 Same-session versus Time-lapse

This study used exactly one probe for each gallery image. The gallery sets (FA|LF) are the same in same-session recognition and time-lapse recognition. The probe set for same-session recognition is made up of images (FA|LM) acquired at about the same time (less than one minute difference) as the probe. The probe set for time-lapse recognition is made up of images (FA|LM) acquired in different weeks from when the gallery images were acquired.

We conducted 9 experiments of different time delays for time-lapse recognition and for each there is a corresponding same-session recognition experiment for comparison.

Figure 5 shows the results for visible and IR. For both modalities, the same session recognition outperforms time-lapse recognition significantly. Note that for same-session recognition there is no clear advantage between IR and visible. However, in time-lapse recognition visible generally outperforms IR.

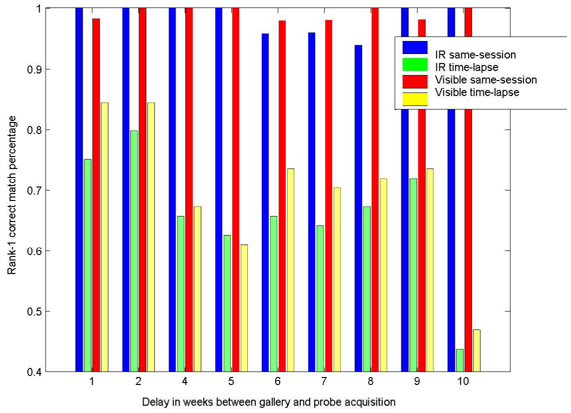


Figure 5: Rank-1 correct match rate for same-session recognition and time-lapse recognition in IR and Visible

7 Sensitivity to Eye Center Location

We manually located eye centers in visible and IR images for normalization. It is possible that error in eye center location could affect the recognition performance differently in visible and IR, especially considering that the IR imagery is more vague than visible imagery and the original resolution for IR is 312 x 219 versus 1600x1200 for visible image. This is potentially an important issue when comparing the performance of IR and visible imagery.

We did a random replacement of the current manually-marked eye centers by another point in a 3x3 (pixel) window, which is centered at the manually-marked position. This is very close to the possible human error in reality. The time-lapse recognition results by using images normalized with the randomly perturbed eye centers are shown in Table 3.

Compared to Table 2, IR is very sensitive to eye center locations. The correct recognition rates drop significantly compared to the performance where the manually located eye centers are used. For visible imagery in time-lapse recognition, the performance decrease is at most slight. This suggests that marking eye centers in IR might be harder to do accurately than marking eye centers in visible, and that this might have affected IR accuracy relative to visible accuracy in our experiments.

Table 3: Rank 1 correct match percentage for time-lapse recognition of combining IR and visible. Top: rank based strategy; Bottom: score based strategy. Row indicates gallery and column indicates probe, eye center is randomly replaced by a point in a 3x3 window that is centered at the manually-located eye center

	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.67 (52) 0.90 (46)	0.65 (44) 0.91 (54)	0.62 (58) 0.71 (55)	0.57 (59) 0.71 (54)
FA|LF	0.68 (40) 0.91 (50)	0.69 (56) 0.92 (27)	0.60 (55) 0.74 (33)	0.62 (61) 0.72 (44)
FB|LM	0.64 (61) 0.75 (56)	0.67 (60) 0.81 (45)	0.65 (62) 0.86 (49)	0.69 (57) 0.84 (50)
FB|LF	0.63 (57) 0.74 (51)	0.62 (57) 0.78 (40)	0.63 (62) 0.88 (33)	0.65 (55) 0.89 (47)

8 Combination of Visible and IR

Table 2 shows that visible imagery is better than IR in time-lapsed recognition, but the sets of mismatched probes of the two classifiers do not necessarily overlap. This suggests that these two modalities potentially offer complementary information about the probe to be identified, which could improve the performance. Since these classifiers yield decision rankings as results, we first consider fusion on the decision level. Kittler *et al.* [12] conclude that the combination rule developed under the most restrictive assumptions, the sum rule, outperformed other classifier combination schemes and so we have used the sum rule for combination in our experiments.

We first used an unweighted rank based strategy for combination.This approach is to compute the sum of the rank for every gallery image. The gallery image with the lowest rank sum will be the first choice of the combination classifier. However, on average, for each probe there are 10-20 rank sum ties (64 gallery images). Since the visible imagery is more reliable based on our experiments in the context of time-lapse, we use the rank of the visible imagery to break the tie. The top of each item in Table 4 shows the combination results using this approach. Only in 2 out of 16 instances is the visible alone slightly better than the combination. The combination classifier outperforms IR and visible in all the other cases.

For each individual classifier (IR or visible), the rank at which all probes are correctly identified is far before rank 64 (64 gallery images). Hence, the first several ranks are more useful than the later ranks. We logarithmically transformed the ranks before combination to put strong emphasis on the first ranks and have the later ranks have a quickly decreasing influence. The middle of each item in Table 4 shows the results of this approach. The combiner outperforms visible and IR in all the sub-experiments and is better than the combiner without rank transformation.

Second, we implemented a score based strategy. We use the distance between the gallery and probe in the face space as the score, which provides the combiner with some additional information that is not available in the rank based method. It is necessary to transform the distances to make them comparable since we used two different face spaces for IR and visible. We used linear transformation, which maps a score s in a range of $I_s = [smin, smax]$ to a target range of $I_{s'} = [0, 100]$. Then we compute the sum of the transformed distances for each gallery and the one with the smallest sum of distances will be the first match. The bottom entry of each item in Table 4 shows the results. The score based strategy outperforms the rank based strategy and improves the performance significantly compared with either of the individual classifiers (IR and visible). This shows that it is desirable to have knowledge about the distribution of the distances and the discrimination ability based on the distance for each individual classifier (IR or visible). This allows us to change the distribution of the scores meaningfully by transforming the distances before combination. This combination strategy is similar to that used by Chang *et al.* [13] in a study of 2D and 3D face recognition.

9 Comparison of PCA and FaceIt

FaceIt is a commercial face-recognition algorithm that performed well in the 2002 Face Recognition Vendor Test[14]. We use FaceIt results to illustrate the importance of combined IR-plus-visible face recognition.

Figure 6 shows the CMC curves for a time-lapse recognition with FA|LF images in the first session as the gallery set and FB|LM images in the second to the tenth sessions as the probe set by FaceIt and PCA. Note that the fusion method is score-based as discussed above. We notice that FaceIt outperforms PCA in visible imagery and IR individually. However, the fusion of IR and visible can easily outperforms either modality alone by PCA or FaceIt. We should take into account the training set PCA used when making this comparison. Given an extremely large unbiased training set which is not often practical

Table 4: Rank 1 correct match percentage for time-lapse recognition of combining IR and visible. Top: simple rank based strategy; Middle: rank based strategy with rank transformation; Bottom: score based strategy. Row indicates gallery and column indicates probe.

	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.91 (25) 0.93 (26) 0.95 (24)	0.95 (23) 0.96 (24) 0.97 (21)	0.83 (45) 0.85 (47) 0.90 (46)	0.81 (44) 0.85 (47) 0.90 (45)
FA|LF	0.91 (18) 0.92 (24) 0.95 (20)	0.93 (19) 0.94 (27) 0.97 (20)	0.85 (41) 0.87 (44) 0.91 (39)	0.83 (23) 0.84 (35) 0.90 (24)
FB|LM	0.87 (20) 0.88 (22) 0.91 (27)	0.92 (34) 0.92 (40) 0.94 (32)	0.85 (23) 0.87 (32) 0.92 (25)	0.86 (32) 0.88 (32) 0.92 (31)
FB|LF	0.85 (43) 0.87 (33) 0.87 (40)	0.87 (40) 0.88 (37) 0.91 (44)	0.88 (12) 0.90 (17) 0.93 (20)	0.90 (36) 0.91 (38) 0.95 (37)

or efficient, PCA might eventually outperform FaceIt in visible-light imagery.

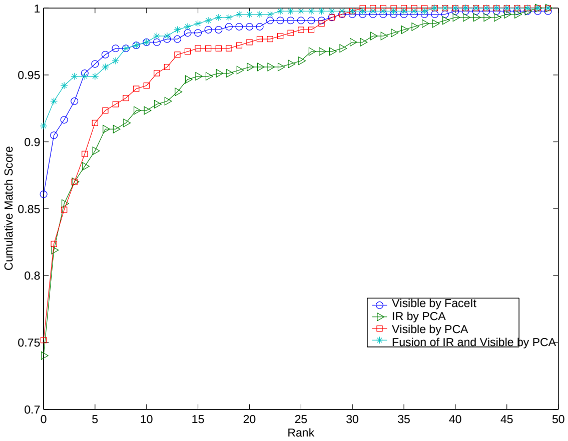


Figure 6: CMC curves of time-lapse recognition using PCA and FaceIt in visible-light and IR

10 Eigenvector Tuning

For one time-lapse recognition with FA|LF images in the first session as the gallery set and FA|LF images in the second to the tenth sessions as the probe set, we examined the eigenvector selection results for IR and visible images.

For IR, we find that dropping any of the first 10 eigenvectors will degrade the performance. A possible reason is that in IR face images, there is no significant unrelevant variance like the lighting in visible images and the first eigenvectors can well describe the true variance between images. When retaining 94% of eigenvectors by removing the last eigenvectors, the performance reaches maximum performance of 82.8%, compared with 81.2% when all eigenvectors are retained. This shows that these last eigenvectors encode noise and are inefficient.

For visible-light, dropping the first 2 eigenvectors make the performance grow to a peak performance of 92.6% from 91.4%. It is possible that some significant unrelevant variance, like lighting, is encoded in these eigenvectors. With these two eigenvectors dropped, We find that retaining about 80% of the eigenvectors by removing the last eigenvectors makes the performance increase to 94.4%, which shows that these last eigenvectors are redundant and undermine the performance.

11 Assessment of Time Dependency

The first experiment is designed to reveal any obvious effect of elapsed time between gallery and probe acquisition on performance. The experiment consists of nine sub-experiments. The gallery set is FA|LF images of session 1. Each of the probes was a set of FA|LF images taken within a single session after session 1 (i.e. sub-experiment 1 used session 2 images in its probes, sub-experiment 2 used session 3 for its probes, and so forth). Figure 7 shows the histogram of the nine rank-1 correct match rates for the nine sub-experiments in IR and visible imagery. The figure shows differences in performance from week to week, but there is no clearly discernible trend over time in the results. All the rank 1 correct match rates in visible imagery are higher than in IR.

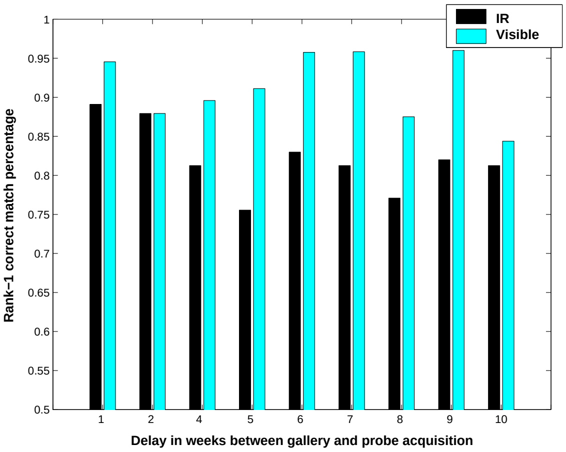


Figure 7: Rank-1 correct match rate for 10 different delays between gallery and probe acquisition in visible and IR

The second experiment was designed to examine the performance of the face recognition system with a constant delay of one week between gallery and probe acquisitions. It consists of nine sub-experiments: the first used images from session 1 as a gallery and session 2 as probe, the second used session 2 as gallery and session 3 as probe and so on. All images were FA|LF. The rank 1 correct match rates for this batch of experiments appear in Figure 8. We note an overall higher level of performance with one week of time lapse than with larger amounts of time. The visible imagery outperforms IR in 7 of the 8 sub-experiments.

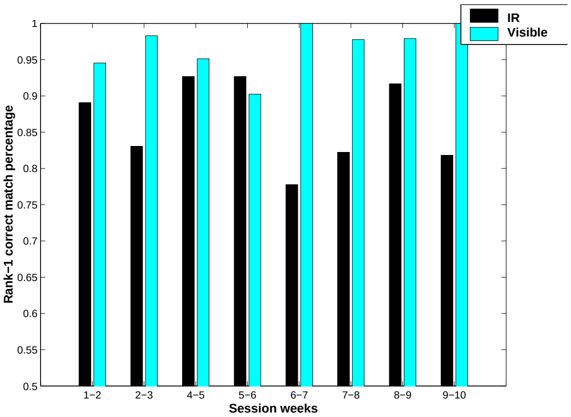


Figure 8: Rank-1 correct match rate for experiments with gallery and probe separated by one week in visible and IR

Together with the time-lapse recognition experiment in Section 7, these experiments show that delay between acquisition of gallery and probe images causes recognition performance to degrade. The one overall surprising result from these experiments is that visible imagery outperforms IR in the context of time-lapse.

11 Statistical Test on Conditions

In Table 2, the probe pairs that are of the same facial expression (lighting condition) but different lighting condition (facial expression), given a gallery of the same facial expression (lighting condition), should reveal the illumination (facial expression) impact. Essentially, we make a comparison of the response of matched pairs of subjects, using dichotomous scales, i.e. subjects are grouped into only two categories, correct/incorrect match at rank 1. Hence we choose McNemar's test [15].

11.1 Illumination Impact

Given the null hypothesis being *there is no difference in performance based on whether the lighting condition for the probe image acquisition is matched to the lighting condition for the gallery image acquisition*, the corresponding $p-$values are reported in Table 5. For IR, what we observed is very likely if the null hypothesis were true and the association between FERET and mugshot lighting conditions for the probe images is NOT significant. However, surprisingly, for visible imagery, there is no evidence to reject the hypothesis either. One reason is that the variance, which is dependent on elapsed-time, dominated over the lighting variance. Another possible reason is that there is not enough difference between FERET and mugshot lighting conditions to produce a noticeable effect. Referring to the images in Figure 1, this explanation seems plausible.

Table 5: p-values of McNemar's test for the impact of lighting change in visible (bottom) and IR (top)

Gallery	Probe pair	p-value
FA|LM	FA|LM	0.55
	FA|LF	0.18
FA|LF	FA|LM	0.50
	FA|LF	0.85
FB|LM	FB|LM	0.50
	FB|LF	0.32
FB|LF	FB|LM	0.51
	FB|LF	0.47

11.2 Facial Expression Impact

Given the null hypothesis being *there is no difference in performance based on whether the facial expression for the probe image acquisition is matched to the facial expression for the gallery image acquisition*, the corresponding $p-$values are reported in Table 6. For visible imagery, all $p-$values are 0, which means that the null hypothesis is unlikely to be true according to what we observed, i.e. the performance is highly dependent on whether the facial expression for the probe image acquisition is matched to the facial expression for the gallery image acquisition. For IR in the group which used neutral expression as gallery, we have the same conclusion as the visible imagery. But for IR with a smiling expression as gallery, we failed to reject the hypothesis, which means the expression impact may be significant in this scenario.

Table 6: p-values of McNemar's test for the impact of expression change in visible (bottom) and IR (top)

Gallery	Probe pair	p-value
FA|LM	FA|LM	0.01
	FB|LM	0.00
FA|LF	FA|LF	0.00
	FB|LF	0.00
FB|LM	FB|LM	0.23
	FA|LM	0.00
FB|LF	FB|LF	0.92
	FA|LF	0.00

12 Conclusion and Discussion

In same session recognition, neither modality is clearly significantly better than another. In time-lapse recognition, the correct match rate at rank 1 decreased for both visible and IR. In general, delay between acquisition of gallery and probe images causes recognition system performance to degrade noticeably relative to same-session recognition. More than one week's delay yielded poorer performance than a single week's delay. However, there is no clear trend, based on the data in this study, that relates the size of the delay to the performance decrease. A longer-term study may reveal a clearer relationship. In this regard, see the results of the Face Recognition Vendor Test 2002 [14].

In time-lapse recognition experiments, we found that: (1) PCA-based recognition using visible images performed better than PCA-based recognition using IR images, (2) FaceIt-based recognition using visible images outperformed either PCA-based recognition on visible or PCA-based recognition on IR, and (3) the combination of PCA-based recognition on visible and PCA-based recognition on IR outperformed FaceIt on visible images. This shows that, even using a standard public-domain recognition engine, multi-modal IR and visible recognition has the potential to improve performance over the current commercially available state of the art.

Perhaps the most interesting conclusion suggested by our experimental results is that visible imagery outperforms IR imagery when the probe image is acquired at a substantial time lapse from the gallery image. This is a distinct difference between our results and those of others [1] [2] [3], in the context of gallery and probe images acquired at nearly the same time. The issue of variability in IR imagery over time certainly deserves additional study. This is especially important because most experimental results reported in the literature are closer to a same-session scenario than a time-lapse scenario, yet a time-lapse scenario may be more relevant to most imagined applications.

Our experimental results also show that the combination of IR plus visible can outperform either IR or visible alone. We find that a combination method that considers the distance values performs better than one that only considers ranks. The image data sets used in this research will eventually be available to other researchers as part of the Human ID database. See *http://www.nd.edu/˜cvrl* for additional information.

Acknowledgments

This work is supported by the DARPA Human ID program through ONR N000140210410 and the National Science Foundation and Central Intelligence Agency through NSF EIA 01-30839.

References

[1] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener, "Comparison of visible and infrared imagery for face recognition," in *2nd International Conference on Automatic Face and Gesture Recognition,Killington,VT*, pp. 182–187, 1996.

[2] D. A. Socolinsky and A. Selinger, "A comparative analysis of face recognition performance with visible and thermal infrared imagery," in *International Conference on Pattern Recognition*, pp. IV: 217–222, August 2002.

[3] A. Selinger and D. A. Socolinsky, "Appearance-based facial recognition using visible and thermal imagery:a comparative study," *Technical Report,Equinox corporation*, 2001.

[4] B. Abidi, "Performance comparison of visual and thermal signatures for face recognition," in *The Biometric Consortium Conference*, 2003.

[5] Y. Yoshitomi, T. Miyaura, S. Tomita, and S. Kimura, "Face identification using thermal image processing," in *IEEE International Workshop on Robot and Human Communication*, pp. 374–379, 1997.

[6] Y. Yoshitomi, N. Miyawaki, S. Tomita, and S. Kimura, "Facial expression recognition using thermal image processing and neural network," in *IEEE International Workshop on Robot and Human Communication*, pp. 380–385, 1997.

[7] *http://www.indentix.com*.

[8] X. Chen, P. Flynn, and K. Bowyer, "Pca-based face recognition in infrared imagery: Baseline and comparative studies," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 127–134, 2003.

[9] *http://www.equinoxsensors.com/products/HID.html*.

[10] *http://www.cs.colostate.edu/evalfacerec/*.

[11] I. Pavlidis, J. Levine, and P. Baukol, "Thermal imaging for anxiety detection," in *IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, pp. 104–109, 2000.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1992.

[13] K. Chang, K. Bowyer, and P. Flynn, "Multi-modal 2d and 3d biometrics for face recognition," *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 187–194, 2003.

[14] *http://www.frvt2002.org*.

[15] M. Bland, *An Introduction to Medical Statistics*. Oxford University Press, 1995.

A Multiscale Approach Weighted with Coherence for Local Frequency Estimation

Kyoungtaek Choi , Sanghoon Lee
and Jaihie Kim
Department of Electrical and Electronic Engineering,Yonsei University
Biometrics Engineering Research Center, Seoul, Korea
Phone: +82 2123–4537
maninquestion@hanmail.net

***Abstract*—This paper suggests a method to estimate the local frequency in a digital noisy image. In order to reduce the effect of noise and estimate local frequency accurately, we use multi-scale estimation. In each scale, we measure the quantity of noise, and along with coherence and we estimate frequency through use of a Kalman filter whose propagation weight is the ratio of coherence between each scales. To demonstrate the performance of our algorithm, we use a noisy simulated image and a fingerprint image.**

***Index Terms*—frequency, multi-scale, Kalman filter, coherence, fingerprint**

I. Introduction

Estimating the local frequency in a digital image is necessary in the overall image enhancement procedure or for use in feature extraction. Especially, the ridge pattern frequency varies not only with indivisual users but also locally due to the pressure which occurs in the capturing process. Therefore in order to filter a ridge pattern or for the distortion caused by pressure, we need to know the local frequency of a given image.

There are several specific approaches used for estimating the local frequency. One is based on spectrum analysis[1], another is the wavelet approach[2] and finally use a Gabor filter[3]. However those methods require a large amount of computation time, so Maio et al. model ridge patterns, sinusoidal signals and frequency estimation in spatial domain uses the partial derivatives of the individual ridge patterns[4]. However, if there is some noise in the image, the algorithm has difficulty estimating the frequency accurately because of a trade-off relationship between the accuracy of the estimated frequency and the reduction of noise. If the algorithm smooths the estimated frequency significantly to reduce the noise, it is difficult to compute the local frequency accurately, while if the algorithm does not smooth the estimated frequency enough to identify the local frequency correctly, it suffers from the noise induced distortion or errors. The uncertainty rule represents this relationship well[5]. To reduce the effect of noise and also estimate the directional information of an image correctly, XiaoGuang et al. suggests a multi-scale orientation estimate method based on PCA (Principal Component Analysis)[6]. This paper combines Maio's and XiaoGuang's concepts properly and suggests the local frequency estimation method which is robust against the spurious noise of an image.

The remaining sections of this paper are organized as follows. Section 2 describes the frequency estimation method applied within a spatial domain. Section 3 describes a multi-scale approach. The experimental results are shown in Section 4. Finally, section 5 contains the conclusion section.

II. Frequency Estimation within a Spatial Domain

Maio et al. modeled a 1D siganl as a sinusoidal signal presented as $f_{\alpha,v}(x) = \alpha \sin(v \cdot x)$. The frequency v of signal $f_{\alpha,v}(x)$ can be calculated through Eq.(1) and (2). If $n\nu$ belongs to a natural number,
$\Gamma^g(f_{\alpha,\nu})$ may equal $\alpha \cdot \nu^g$.

$$\Gamma^g(f_{\alpha,\nu}) = \frac{1}{n}\int_0^{n\frac{\pi}{2}} \left|\frac{d^g f_{\alpha,v}(x)}{dx^g}\right| dx = \alpha \cdot \nu^g, \;\; (nv) \in Z^+ \quad (1)$$

$$v = \frac{\Gamma^{g+1}(f_{\alpha,v})}{\Gamma^g(f_{\alpha,v})}, \text{g} = 0, 1, ..., \infty \quad (2)$$

When n approaches to ∞, the requirement of Eq.(1) can be satisfied[4]. This 1D signal $f_{\alpha,v}(x)$ can be expanded to a 2D signal $f_{\rho,\theta,\alpha,v}(x,y) = \rho + \alpha \sin(v(x \sin\theta + y \cos\theta))$. In 2 dimensions, Eq.(2) becomes Eq.(3).

$$v = \sqrt{\frac{(\Gamma_x^2)^2 + (\Gamma_{xy}^2)^2 + (\Gamma_{yx}^2)^2 + (\Gamma_y^2)^2}{(\Gamma_x'^1)^2 + (\Gamma_y'^1)^2}} \quad (3)$$

$$\Gamma_x^2 = \frac{2}{\pi \cdot n^2} \int_0^{n\cdot\frac{\pi}{2}} \int_0^{n\cdot\frac{\pi}{2}} \left|\frac{\partial^2 f_{\rho,\theta,\alpha,v}(x,y)}{\partial x^2}\right| dxdy \quad (4)$$

$$\Gamma_{xy}^2 = \frac{2}{\pi \cdot n^2} \int_0^{n\cdot\frac{\pi}{2}} \int_0^{n\cdot\frac{\pi}{2}} \left|\frac{\partial f_{\rho,\theta,\alpha,v}(x,y)}{\partial x \partial y}\right| dxdy \quad (5)$$

$$\Gamma_x'^1 = \sqrt{\frac{8}{\pi^2 n^2} \int_0^{n\cdot\frac{\pi}{2}} \int_0^{n\cdot\frac{\pi}{2}} \left(\frac{\partial f_{\rho,\theta,\alpha,v}(x,y)}{\partial x}\right)^2 dxdy} \quad (6)$$

In the 2D domain, the 1st derivative and the 2nd derivative can be calculated through the use of a gradient operator and Hessian matrix[7]. To estimate frequency accurately in a noisy image, n must be large, but if n is too large, the local variety of the available frequency can disappear. If we decrease the size of the window, we can estimate the local frequency more precisely but accuracy suffers from the noise of the image. We use the term large, because no specific measure is available, so large is the subjective term used in this paper. Maio et al. can not suggest a reliable method to reduce the noise and measure the local variety of the frequencies simultaneously, so in the next section, we suggest multi-scale frequency estimation method to adjust for the trade-off between noise reduction and frequency localization.

III. Multiscale Frequency Estimation

To estimate local frequency in a noisy image, we need to measure the noise and adaptively adjust the frequency estimation method to the existing noise. There are some theses proposed to measure noise quantities[8][9]. Among the several noise measures, Coherence is a well-formatted and good measure used in determining the ridge pattern of an image, so we select the coherence as a measure of noise. Coherence is defined as the ratio between the difference of the maximum and minimum eigen-value and the summation of these values as presented in Eq.(7). The symbols $\lambda_{\max}$ and $\lambda_{\min}$ are maximum and minimum eigen-value repectively. To calculate coherence, we divide an image into several blocks and calculate the covariance matrix of the image gradient vector in each block. After completing the process, we can calculate the eigen-vectors and the eigen-values of the matrix through use of the Singular Value Decomposition (SVD)[10]. Instead of using the SVD, Asker et al. suggest a direct calculation method to caculate the eigen-vectors and the eigen-values simply[11]. Coherence R can be expressed another way as shown in Eq.(8).

$$R = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \tag{7}$$

$$R = \sqrt{\frac{(G_{xx} - G_{yy})^2 + 4G_{xy}^2}{G_{xx} + G_{yy}}} \tag{8}$$

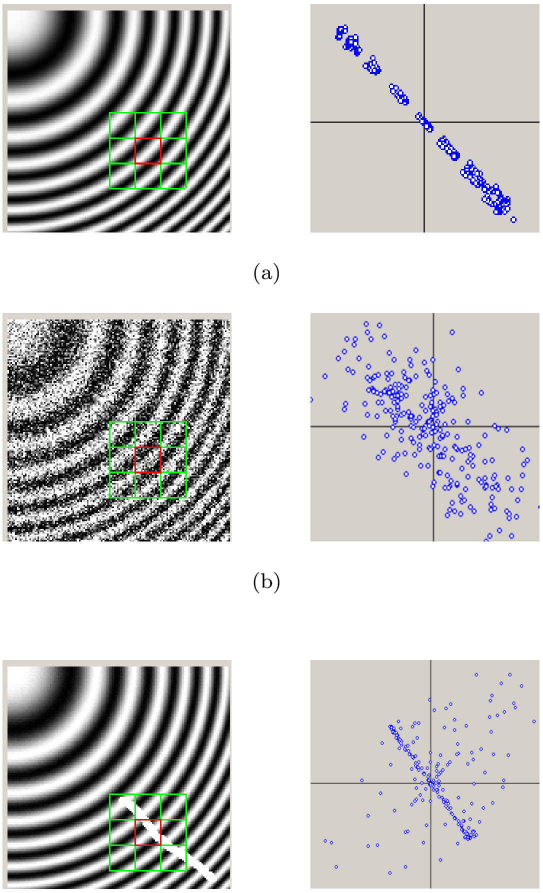
where

$$G_{xx} = \sum_N G_x^2 \tag{9}$$

$$G_{yy} = \sum_N G_y^2 \tag{10}$$

$$G_{xy} = \sum_N G_x G_y \tag{11}$$

In Eq.(11) G_x and G_y are the x and y elements of the gradient vector used in the Cartesian coordinate and N is the window size. In a ridge pattern image, like a fingerprint, coherence can indicate how uniform the directional information is [6]. If there is some white Gaussian noise in the individual ridge patterns, the distribution of the gredient vectors disperses, otherwise the distribution has uniform direction as displayed in Fig.1. The size of the window in which the gradient vectors are calculated, is 12 × 12 pixels. Scars or scratches are different from white Gaussian noise but the distribution of the gradient vectors also disperses as shown in Fig.1(c). Since coherence is defined as the normal difference between maximum eigen-vale and minimum eigen-value, if the gradient vectors distribute widely, coherence is reduced.



(a)

(b)

(c)

Fig. 1. the distribution of gradient : (a) non-noise image, (b) gaussian noise image (c) scar noise image

If the noise is white Gaussian, the PDF of R is determined as presented in Eq.(12)[6]. According to window size of N, the PDF is shown in Fig.2.

$$p(R) = 4(N-1)R\frac{(1-R^2)^{N-2}}{(1+R^2)^N} \tag{12}$$

If we presuppose that as R is decreasing, the probability of the existance of noise increases. so we can use R as the measure of noise quantity.

To estimate the local frequency, we use a multi-scale approach, using R as the propagation weight. If we convolve

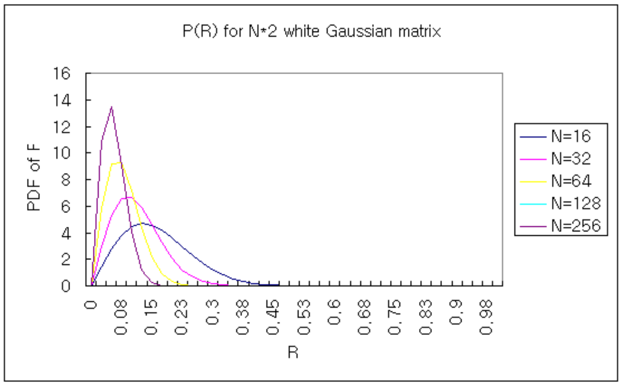


Fig. 2. PDF of R(Coherence) for white Gaussian noise

the image with a low-pass filter and downsample it to focus on low levels of resolution, we can reduce the noise but, we can also distort the frequency information too. Therefore instead of downsampling the image, we adjust the window-size n for several layers as previously presented in Eq.(4), Eq.(5), Eq.(6). When we calculate Γ_x^2, in order to exclude the noisy region we use Eq.(13) instead of Eq.(4). We change Eq.(5) and Eq.(6) to Eq.(14) and Eq.(15) respectively. In Eq.(13) $R(x,y)$ is coherence in the x,y position and the threshold th is defined by referring to Eq.(12).

$$\Gamma_x^2 = \frac{2}{\pi \cdot (\sum\limits_n \sum\limits_n T_{(x,y)})} \sum_n \sum_n T_{(x,y)} \cdot \left| \frac{\partial^2 f(x,y)}{\partial x^2} \right| \tag{13}$$

,where if $R(x,y) > th$ $T_{(x,y)} = 1$ else $T_{(x,y)} = 0$

$$\Gamma_{xy} = \frac{2}{\pi \cdot (\sum\limits_n \sum\limits_n T_{(x,y)})} \sum_n \sum_n T_{(x,y)} \cdot \left| \frac{\partial f(x,y)}{\partial x \partial y} \right| \tag{14}$$

$$\Gamma'^{1}_{\mathrm{x}} = \sqrt{\frac{8}{(\pi \sum\limits_{\mathrm{n}} \sum\limits_{\mathrm{n}} \mathrm{T}_{(\mathrm{x,y})})^2} \sum_{\mathrm{n}} \sum_{\mathrm{n}} \mathrm{T}_{(\mathrm{x,y})} \times \left(\frac{\partial \mathrm{f(x,y)}}{\partial \mathrm{x}} \right)^2} \tag{15}$$

According to the layer selected, increasing the window size makes the algorithm reguire a larger amount of calculations. When the window size is large, the windows overlap each other to the extent that the estimated frequency is little different from those of the neighbor blocks and it can enlarge both the opportunity for errors as well as redundancy. Since we estimate the frequency not in an individual pixel but in an entire block, we have to overlap windows to reduce the block effect. However, in order to reduce redundancy we can adjust the overlapped region by decreasing the amount of frequency estimation within each layer as presented in Fig.3. In Fig.3 the dots(frequency measure point) are the center of the frequency measure window which is colored gray. As the layer goes up to higher level, the number of dots is increasing and the size of the window is decreasing. We measure the local frequency around each dot in the frequency measure window of each layer and average the coherence which is already calculated in the 16×16 sized window. We can designate the frequency measured in each layer the observation of the layer. We can estimate the frequency in the current layer with the observation of the current layer and the estimation of the previous layer by using a Kalman filter as explained in Eq.(16)[6].

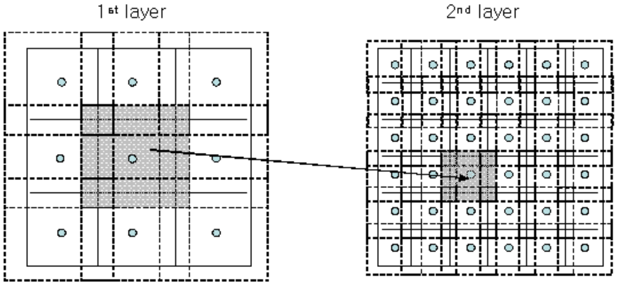


Fig. 3. Frequency Sample Point in each layer

$$\hat{s}[n] = \hat{s}[\gamma n] + \frac{R_n}{R_n + R_{\gamma n}} (x[n] - \hat{s}[\gamma n]) \tag{16}$$

where $x[n]$ and $\hat{s}[n]$ is the observation and the frequency estimation of the current layer repectively. The symbol $\hat{s}[\gamma n]$ represents the frequency estimation in the previous layer. The symbols R_n and $R_{\gamma n}$ are the coherence of the current layer and the previous layer respectively.

IV. Experimental Results

We estimate the local frequency through 4 layers and the sizes of the frequency measure windows in each layer are 129×129, 65×65, 33×33 and 17×17. The number of frequency measure points in each layer is a quarter of the child layer's of each corresponding layer. Coherence is calculated once in the 16×16 sized window. First to show how well our algorithm estimates the local frequency, we excuted an experimental test with a fan-shaped ridge pattern Fig.4. The frequency of Fig.4(a) is linearly increasing according to the distance from the point of origin. Fig.4(b) shows the frequency of Fig.4(a) as a grayscale value. To prove that our method can estimate the local frequency well, we filter the test image through a constant frequency Gabor filter and an adaptive Gabor filter. Fig.4(c) represented the image filtered with a constant frequency Gabor filter whose frequency is 0.13 and Fig.4(d) represented the image filtered with an adaptive Gabor filter which varies its frequency from 0.11 to 0.17, according to the estimated frequency. Since our method estimates the local frequency successfully, as shown in Fig.4(d), ridge patterns are welll-extracted. Because of the limits of low and high filter frequencies, there is some distortion as shown in Fig.4(d).

Fig.5 shows that our algorithm is more robust than Maio's in estimating the frequency of an image. In Fig.5(b) which represents the grayscale value of unsmoothed frequencies of Fig.5(a), we determine that Maio's algorithm is very sensitive to noise. To reduce the noise, we smoothed the frequency with 7×7 blocks average filter(56×56 pixel)

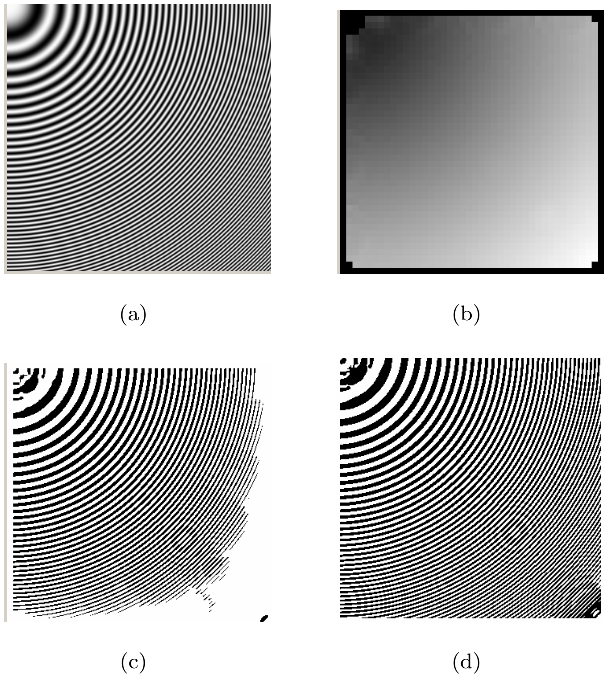


Fig. 4. local frequency : (a) test image, (b) multi-scale frequency, (c) constant gabor filtering, (d) adaptive gabor filtering

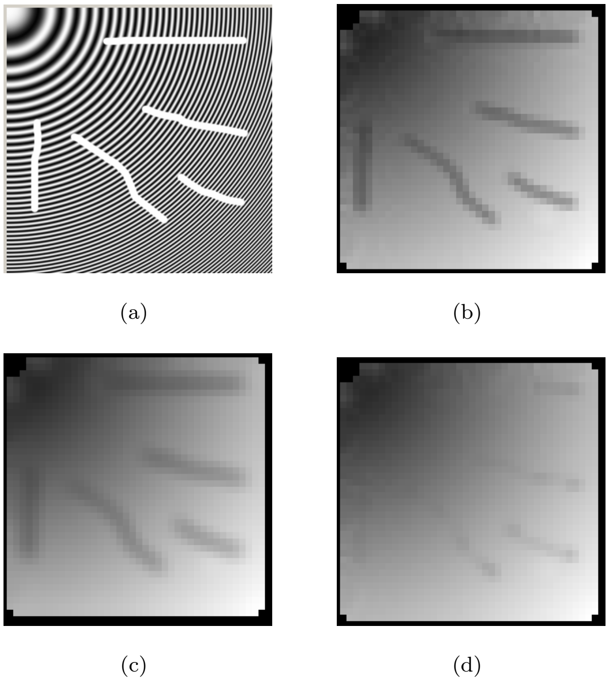


Fig. 5. local frequency of scar image : (a) scar image, (b) non-smoothing frequency, (c) 7×7 smoothing frequency, (d) multi-scale frequency

as presented in Fig.5(c). However, we can see the effect of noise near the scar and if we increased the filter size, we lost the local frequency information. Fig.5(d) represented the frequency image estimated by our algorithm. Our algorithm reduces the noise effectively and does not distort the local frequency. Table.I summarizes the comparison of our algorithm and Maio's algorithm as the average filter size increases. Since we calcaulte the block frequency, the unit of measure for filter size is the number of blocks where one block occupies 8×8 pixels. First, we regard the non-smoothing frequency of Fig.4(a) as a reference, which has a correct frequency and we caculate error as shown in Eq.(17). The $Ref(x, y)$ is the reference and $F(x, y)$ is the smoothed frequency image or the multi-scale frequency image. We predetermined that there is no problem in regarding the non-smoothing frequency of Fig.4(a) as the reference frequency because Fig.4(a) has no noise. From Table.I, we can see that simple frequency smoothing dose not solve the trade-off problem between noise reduction and local frequency accuracy, but our multi-scale method can obtain the optimum solution and also it has almost minimum error observed in both cases (noisy image and non-noisy image).

$$Error = \sum \sum (F(x, y) - \mathrm{Ref}(x, y))^2 \qquad (17)$$

We use our algorithm to estimate the frequency of a fingerprint image as shown in Fig.6. In Fig.6, the interior region of the red circle is a higher frequency region than other frequenciess of the fingerprint. We can see that our algorithm indicates the high frequency region with higher level of grayscale brightness than other regions and it can estimate the local frequency of real image well too. We filtered the fingerprint image with a simple Gabor filter and an adaptive Gabor filter to extract ridge patterns. After filtering the image with a simple Gabor filter, which has constant frequency parameter of 0.13, the ridge patterns of the interior red circle are broken as displayed in Fig.6(c), but after adaptively using Gabor filter, the ridge patterns are very similar to the patterns of the original image as shown in Fig.6(d).

TABLE I
Frequency Square Error

Filter	4×4	8×8	12×12	*multiscale*
$Fig.4(a)$	0.085	0.168	0.292	0.063
$Fig.5(a)$	6.06	4.08	3.21	0.792

Unfortunately, we did not have an adequate number of fingerprint images whose local frequency varies greatly, so we could not complete the verification test by applying different Gabor filters.

V. Conclusion

This paper proposed a local frequency estimation method by using a multi-scale approach. To estimate the local frequency, we used Maio's algorithm because it required less computation than others and has easily integrated with the multi-scale approach. To reduce the effect of noise, first we measured the quantity of noise with the coherence of the image and applied a multi-scale approach to Maio's algorithm, weighted with coherence. In the ex-

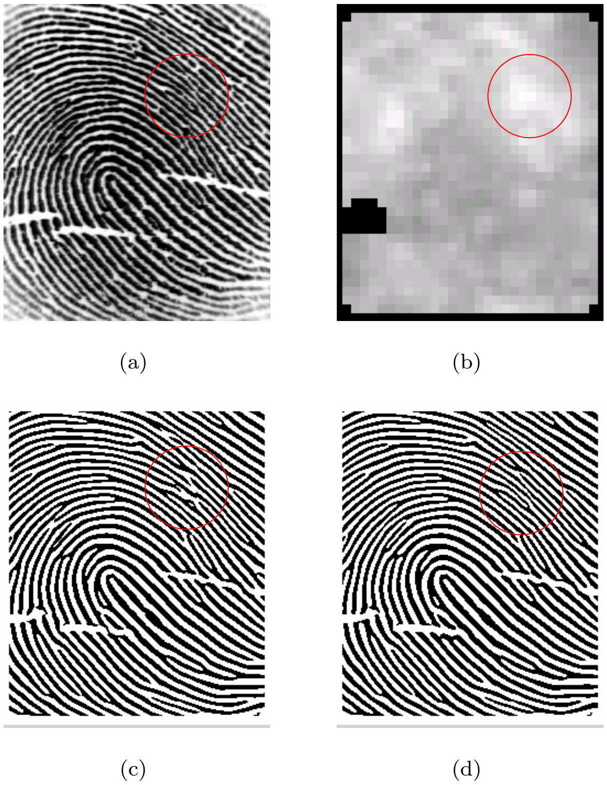

(a)
(b)
(c)
(d)

Fig. 6. local frequency of scar image : (a) fingerprint, (b) multi-scale frequency, (c) gabor filtered imag, (d) adaptive gabor filtered image

perimental result, the multi-scale approach was proven to be more accurate and robust than simply smoothing algorithm in both a non-noise image and noisy image. We used this algorithm to adjust the filter parameter in the extraction of the ridge pattern of an individual fingerprint image or to compensate for the distortion of a fingerprint caused by pressure on the capture device. We will use this algorithm to solve the latter problem by making the ridge intervals the same through the entire warpping process.

Acknowledgements

This work was supported (in part) by Biometrics Engineering Research Center,(KOSEF).

References

[1] Alan V. Oppenheim, Alan S. Willsky, S. Hamid Nawab, “Signals and systems,“ *Prentice Hall*, 1997.

[2] S. Mallat , “a wavelet tour of signal processing” *Academic Press*, 1998.

[3] R. Buse, Z.Q. Liu, and T. Caelli, “Using Gabor filters to measure the physical parameters of lines,” *Pattern Recognition*, vol.29, no.4, pp.615-625, 1996. vol. 2, pp. 187–193, 1999.

[4] D. Maio, D. Maltoni, “Ridge-line Density Estimation in Digital Images,“ *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, vol.1, pp.534 -538, 1998

[5] G. Strang, T. Nguyen, “Wavelets and Filter Banks,“ *Wellesley-Cambridge Press*, 1996.

[6] X. G. Feng, P. Milanfar, “Multiscale principal components analysis for image local orientation estimation,” *The 36th Asilomar Conference on Signals, Systems and Computers*, 2002.

[7] R.C. Gonzalez, R.E. Woods, “Digital Image Processing,” *Addison Wesley*, 1992.

[8] J. Weickert, “Coherence-Enhancing Diffusion Filtering,” *International Journal of Computer Vision 31*, pp. 111-127, 1999.

[9] A.M. Bazen, S.H. Gerez “Systematic Methods for the Computation of the Directional Fields and Singular Points of Fingerprints,” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 24, no. 7, Jul , 2002.

[10] T.K.Moon, W.C. Stirling, “Mathematical Methods and Algorithms for Signal Processing,” *Prentice Hall*, 2000.

[11] A.M. Bazen, S.H. Gerez “Directional Field Computation for Fingerprints Based on the Principal Component Analysis of Local Gradients,” *Proc. ProRISC2000, 11th Ann. Workshop Circuits, Systems and Signal Processing*,Nov , 2000.

Speaker Verification with Bayesian Networks

Eduardo Sánchez-Soto, Raphaël Blouet, Gérard Chollet and Marc Sigelle

École Nationale Supérieure des Télécommunications
Département de Traitement de Signal et des Images. LTCI/CNRS URA 820.
46 rue Barrault 75634 Paris Cedex 13 France
esanchez,raphael.blouet,gerard.chollet,marc.sigelle@tsi.enst.fr

Abstract

One solution to improve the performance of Speaker Recognition (SR) systems could be the integration of different aspects of the speech signal. Thus in this paper it is proposed to integrate, or fuse, all these informations in a probabilistic framework with a system based on Bayesian Networks (BNs) where the structure is learned directly from the data. BNs are a flexible and formal statistical framework that allows us to represent the conditional independence relations among different speech features that convey information about the speaker identity. In this paper, prosodic variables (pitch and energy), the linear prediction cepstral coefficients (LPCC) from signal and LPCC from residual signal of linear prediction analysis are used to represent each speaker.

This study is conducted on the NIST 2002 one speaker text-independent data base. These experiments confirm the potentialities of BN approach.

1. Introduction

Speech signal carries a lot of information besides the message. Other information about the speaker is present such as mood, emotive state and in particular his/her identity. SR (Identification (SI) or Verification (SV)) systems should use features which capture characteristics of the speaker in order to differentiate them from others. In this search for individual discriminant features some information could be lost. Many authors discard prosodic information in speaker verification, but it is known that they carry a lot of information about the speaker identity. Therefore speaker information of other sources must be used. The suprasegmental characteristics, like intonation, accent or pitch are really important in a normal communication, specially the pitch that appears like an important factor in speaker recognition [1]. However the pitch information in itself is not enough to discriminate between two different persons. Therefore speaker information of other sources must be used. For example, spectral information, conveyed by cepstral coefficients, and knowledge, which is not often taken in account, that comes from the source of excitation in speech production.

The main idea, developed in this paper, is to retrieve the conditional independencies directly from the data (linear residual analysis from the source in speech production, the spectral information from the vocal tract and prosody) in order to build a BN, and by this mean integrate, in a probabilistic way, all those informations.

This paper is organized as follows: Bayesian Networks are first introduced in section 2, with some discussion about the inference problem and algorithms. Section 3 reviews briefly some ideas about structure and parameters learning in BNs. In section 4, the experiments, results and their probabilistic interpretation are presented. Finally conclusions and perspectives are given in section 5.

2. Bayesian Networks

A BN, or Bayesian Belief Network [2], represents a joint probability distribution defined on a finite set of random variables. It is a formal representation, based on probability theory and graph theory, given by a Directed Acyclic Graph (DAG) in which nodes represent random variables and arcs represent conditional probabilistic dependencies among those variables. An arc from **Q** to **Y** can also be interpreted as indicating that **Q** has a direct influence on **Y**, Figure 1.

In a DAG each edge points from one node, called parent, to another, called child. In the same topology description, the node X_j has a descendant node X_i if this one is its child or is connected to it through its children. In a BN, a conditional probability distribution is associated with each node X_i that describes the dependency between this node and its parents, each node is conditionally independent from its non-descendants given its parents. Those dependence relations induces a factorization in the joint distribution function expressed as :

$$P(X_1,\ldots,X_N)=\prod_{i=1}^{N}P(X_i|Pa(X_i)), \tag{1}$$

where $Pa(X_i)$ is the set of X_i's parents.

2.1. Inference

There are two main research problems in probabilistic reasoning using Bayesian Networks: learning and inference [3]. Bayesian network inference involves computing the posterior marginal probability distribution of some query nodes, and computing the most probable explanation given the values of some observed nodes once the structure is known.

A BN is a couple $(G, CPDs)$ formed by one structure, the graph G, and a set of Conditional Probability Distributions (CPD), one for each node with parents in the network. For nodes without parents we have just to specify their prior probability. Evidence, i.e. knowledge about the state of one variable, would modify the states of others variables in the network. Doing probability inference consists in computing the probability of each state of a node when we know the state taken by some other variables. There are three types of evidence propagation, exact, approximated and symbolic. One or another is used depending on the characteristics of the data and the complexity of the structure. In order to make exact inference it is necessary to talk about ”belief propagation” [4] and to take into account the relation of independence obtained directly from the graph. The exact methods present some problems. Some of them are not applicable to all the types of structures. The methods of general validity become very inefficient with certain structures when the number of nodes and its complexity grow. This is not surprising since it has been demonstrated that the exact propagation task is NP-hard [5]. For that reason, and from a practical point of view, the exact propagation methods can be very restrictive and even inefficient in situations in which the type of structure of the network requires a large memory and a lot of computational power. With the second method, approximated values are obtained using simulation methods as Monte Carlo and Gibbs sampling [6]. The last method of propagation works directly with symbolic parameters [7].

In general, if we have a set of variables $X = \{X_1, X_2, \dots, X_N\}$ and a set E, the evidence, with known values $E = \{e_1, e_2, \dots, e_M\}$, where $E \subset X$, inference consists in computing :

$$p(x_i|e) = \frac{p(x_i, e)}{p(e)} \; \alpha \; p(x_i, e). \tag{2}$$

The conditional dependence assumptions encoded by a BN have the advantage of simplifying the conditional probabilities computation. All this could be done in an equivalent tree structure when the original one is not a tree [8]. This structure is a tree built of cliques that represent the local structures, and then preserve the conditional probabilities. The first step in the junction tree construction consists in finding those cliques C_i. Then it is possible to compute their CPD. The CPDs of variables X_i are computed by marginalizing the cliques. In detail this process works has follows:

1. moralization and triangulation (because the parents are correlated given its children) of G to obtain an undirected graph G'.
2. computation of cliques C of G',
3. assign each X_i from X to one clique C_i,
4. for each $C_i \in C$ define a potential $\psi_i(C_i) = \prod_{X_i \in A_i} P(x_i|Pa(x_i))$.

After those steps, the belief propagation method has to be applied to the new graph (collecting and distribution steps). That is, it must be updated the belief in each node when some variables have been observed.

3. Learning

The other main problem in probabilistic reasoning using Bayesian Networks is learning. Learning Bayesian Network from data [9] [10] consists in automatically constructing the network, structure and parameters, from information in data using some learning algorithms. The Statistical base of BN let the development of learning methods. We use these methods in order to obtain the conditional independences in the graph structure and the conditional probability distributions that quantify those dependences directly from databases. Therefore, dependences, structure and conditional probability distributions can be learned from data.

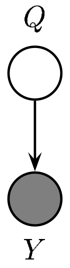


Figure 1: *Basic BN.*

3.1. Structure

In the process of finding the best structure, even if the space of variables is fully observable, some aspects must be considered. Firstly concerning the structure space, should trees be a priority or should more complex graphs be considered? The number of possible structures depends on the number of variables n in a super-exponential way. For example, with four variables there are 543 possible DAGs. It is unrealistic to explore all of them. For that reason, it has to be taken in consideration search algorithms that gives the structures to be evaluated. There are two different approaches to solve this problem, the first one, like MCMC [11], searches in all the structure space and returns either the best one, or the best in a Markov equivalent way. The second approach starts with a specific connected graph and then searches for independence relations in the data S, and puts in or takes away arcs.

The K2 algorithm [12], used in this work, belongs to the second approach. It starts with a structure, the simplest one, i.e. a graph without arcs. It needs some prior knowledge and a relationship between the variables. Then, for each variable X_i we look for the set $Pa(X_i)$. The variables in this set are restricted to those variables with smaller order numbers than X_i.

In order to achieve learning, a scoring function must be specified for measuring the network's quality. The criterion, or quality measure to select $Pa(X_i)$ is the last aspect to study in the structure learning. Maximum likelihood could be an adequate quality measure, but it privileges the fully connected graph. This graph gets the highest likelihood because it has the greatest number of parameters. Thus, to overcome this problem, a prior knowledge on the model can be used. By Bayes' rule, the MAP model is the one that maximizes :

$$P(G|S) = \frac{P(S|G)P(G)}{P(S)}, \tag{3}$$

where $P(G)$ penalizes complex model and $P(S)$ is a constant. The marginal likelihood is :

$$P(S|G) = \int_\theta P(S|G,\theta)P(\theta|G)d\theta, \tag{4}$$

where S is the database. (4) as the advantage that automatically penalizes more complex structures. This score function can be approximated [13] with a Laplace method, and finally get the BIC (Bayesian Information Criterion) :

$$logP(S|G) \approx logP(S|G,\hat{\theta}) - \frac{d}{2}logM, \qquad (5)$$

where M is the number of samples, $\hat{\theta}$ is the ML estimate of the parameters and d is the dimension of the model.

3.2. Parameters

Here, it is required to adjust the parameters of the BN in such a way that the CPDs describe the data statistically. The parameters θ and the model, $B(\theta)$, defined for these parameters are given. Also, the prior distribution over the models $P(B(\theta))$ and the space of parameters in these models $P(\theta|B)$ can be used. So, given some data S, it is wanted to estimate θ, such that the posterior probability to be maximized is :

$$P(B|S) = \frac{P(B)}{P(S)} \int_{\theta} P(S|\theta, B)P(\theta|B)d\theta. \qquad (6)$$

Thus the maximum likelihood estimate of θ is computed by minimizing the cost function over the probability density function. We can make an optimization that relies on the gradient of this function, or use an iterative procedure called Expectation - Maximization (EM) [14] or a variant, Generalized EM , using a gradient method in the M step.

4. Experiments and Results

In this section, experiments and results using our BN Speaker Verification System (BNSVS) are detailed.

4.1. Database

The data are taken from the second release of the Cellular Switchboard Corpus (Switchboard Cellular - Part 2) of the Linguistic Data Consortium (LDC) [15]. Each conversation is echo cancelled before use. The database is divided into training data (about 400 target speakers), and test data (about 3500 test segments). The training data for a target speaker consist in about two minutes of speech from that speaker, excerpted from a single conversation. Actual duration is, however, constrained to lie within the range of 110 to 130 seconds. Each test segment is extracted from a 1 minute excerpt of a single conversation and is the concatenation of all speech from the subject speaker during the excerpt. The duration of the test segment therefore vary, depending on how much the segment speaker spoke. So, the effective speech duration lies between 15 and 45 seconds. Both test and target speakers are of the same sex.

4.2. Modeling

The training and test parameter vectors consist of a set of four types of parameters. The first vector is a 24-dimensional LP Cepstral Coefficients obtained as follow : 12-dimensional LPCC, with sliding CMS (Cepstral Mean Substraction) and augmented with their first derivatives, $SLPCC$, for Signal Linear Prediction Cepstral Coefficients. The second vector, 24-dimensional LP Cepstral Coefficients has been obtained as before from the LP-residual signal $RLPCC$ [16][17], and finally the frame pitch F_0 and the frame energy E.

Those data had been used with K2 algorithm to find the best structure for our four variables. We have worked with all the possible orders and used the BIC score [5]. From this analysis we have obtained the conditional independence relations for the multivariate Gaussian distribution that define the network structure which is set to be speaker independent, Figure 2.

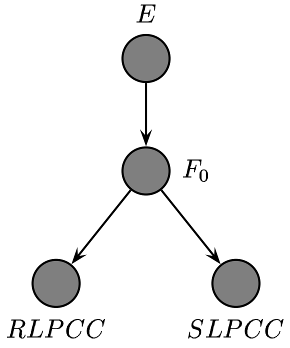


Figure 2: *Structure for the four variables (energy (E), pitch (F_0), signal $SLPCC$ and residual $RLPCC$) issued from the K2 algorithm.*

From basic probability theory the joint probability for the four variables $U = \{E, F_0, RLPCC, SLPCC\}$ can be written as:

$$\begin{aligned} P(U) = P(E)P(F_0|E)P(RLPCC|F_0, E) \\ P(SLPCC|F_0, E, RLPCC). \qquad (7) \end{aligned}$$

Now, taking into account the graph of Figure 2 and its relations of conditional independence, this equation becomes a product of local terms :

$$\begin{aligned} P(U) = P(E)P(F_0|E)P(RLPCC|F_0) \\ P(SLPCC|F_0). \qquad (8) \end{aligned}$$

The relation between $SLPCC$, $RLPCC$ and F_0 is obtained from the term $P(RLPCC|F_0)P(SLPCC|F_0)$. It can be interpreted as a relation of conditional independence where $RLPC$ and $SLPC$ are independent given F_0, noted $RLPCC \perp SLPCC|F_0$ or $I(RLPCC, SLPCC|F_0)$. Also, from the second term in (8) it can be seen that F_0 depends directly of E.

The physical interpretation of the relations between the variables gives the same relations found in the equations obtained from the graph. For example, the voiced speech has more energy that the unvoiced speech. It is evident that the speech energy depends directly from the speech voicing. This fact is written in the term $P(F_0|E)$. The source influences the spectral envelope due to the filtering effect of the vocal tract. The pitch is correlated with the vibration of the vocal folds and the vocal tract characteristics. Consequently, the source and the spectral envelope depends on pitch as it is seen in $P(RLPCC|F_0)P(SLPCC|F_0)$.

The relations obtained in equation (8) exhibit the causal interaction between the variables. Now, using Bayes theorem : $P(E)P(F_0|E) = P(E)P(F_0|E)$, the equation (8) can be rewritten as :

$$\begin{aligned} P(U) = P(F_0)P(E|F_0)P(RLPCC|F_0) \\ P(SLPCC|F_0). \qquad (9) \end{aligned}$$

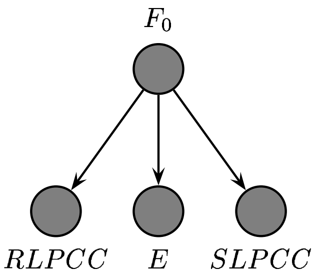


Figure 3: *Equivalent structure for the four variables (energy (E), pitch(F_0), signal $SLPCC$ and residual $RLPCC$) using the equality $P(E)P(F_0|E) = P(E)P(F_0|E)$.*

This new formulation corresponds to the graph shown on Figure 3. In this equation the causal relations represented are not similar to that presented in (8), but the probability density function is the same. Then the equation (9) also represents the variables relation. This structure has the advantage that pitch is the root node. Pitch is a feature whose domain is longer than just one single phonetic segment. Then the independence relations found in the equation (9) represent the conditional independence of $SLPCC$, $RLPCC$ and E given F_0. Where F_0 is a prosodic variable that relate different linguistic elements, by making boundaries and defining transitions in speech signal.

Once the structure has been learned, the final Universal Background Model (UBM) BN's parameters are learned. Since there are not enough training data for each speaker, adaptation methods are applied to compute every Target Speaker Model. For this purpose, the system starts from an universal model (UBM) which is then adapted to the client speaker by three iterations of the GEM algorithm and in this way we overcome the problem. Two gender-dependent UBM have been created using part of the 2001 cellular development and evaluation datasets (this database is similar to the database already described).

4.3. Results

Each test segment is evaluated against 11 hypothesized speakers. The decision score is directly based on the log-likelihood ratio between the target speaker and the UBM over all the frames without any kind of normalization. Figures 6 and 5 display the DET (Detection Error Tradeoff) curves that measure the performance obtained with our system and the standard technique Gaussian Mixture Models (GMM), that have become the dominant approach for modeling multivariate densities in text-independent speaker recognition. A DET curve is a mean of representing performances on detection tasks and is an standard in speaker and language recognition evaluations. In a DET curve, error rates are plotted on both axes (False Alarm and Miss Detection). It shows when a system fails to detect a target or declare such a detection when the target is not present.

First experiment uses the vector $SLPCC$ modelled by a GMM with 64 mixtures. The results shown in the DET curve, Figure 5, show a performance of 19.31 % at the Equal Error Rate (EER). The same has been done with the $RLPCC$ vector obtaining a score of 24.34 %. Now combining all the variables in a vector and using a GMM with 64 mixtures a 21.34 % score is obtained.

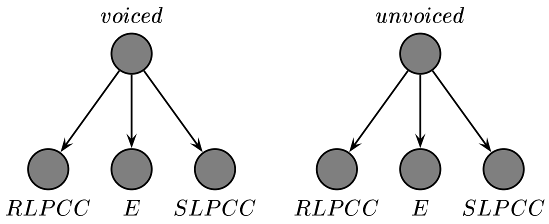


Figure 4: *Structure used in the second experiments.*

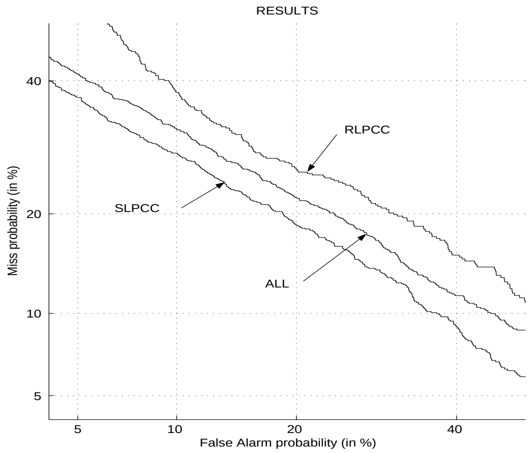


Figure 5: *DET curve for NIST 2002 evaluation data with SLPCC, RLPCC and All using a GMM with 64 mixtures.*

The next set of experiments use two models. The first one uses the structure in the Figure 2 and the set of parameters: 32 Gaussians for $RLPCC$ and $SLPCC$ plus 2 for the pitch F_0 and energy E. CPDs were learned with GEM [18] [19]. This choice of gaussian numbers (parameters number) was made taken into account the computation resources and time requests to finish a task. K-means was used to determine the initial setting for the Gaussian parameters. This system obtains an EER of 24%, Figure 6. The results in the Figures 5 and 6 show that a GMM with a $SLPCC$ vector perform better than our first system. Given that our score is similar to that obtained with the $RLPCC$ vector the difference can come from the independence relations obtained in the structure.

With the second structure shown in the Figure 4, a discretization of the continuous pitch F_0 was made in order to better modelize the voiced and unvoiced parts of speech. The parameters used for this model are : 2 values for the pitch (voiced and unvoiced), 16 Gaussians for the $RLPCC$ and $SLPCC$ and 2 Gaussians for the energy E. This system, shown in Figure 6 obtains an EER of 21.18% for male and 22.37% for female.

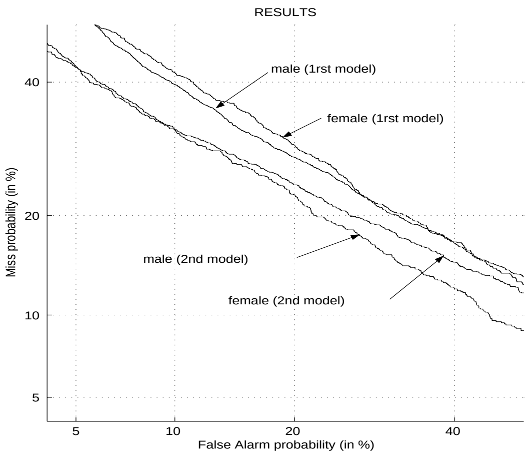


Figure 6: *DET curve for NIST 2002 evaluation data using our two Bayesian Network models: First Model as shown in Fig. 2 and Second Model as shown in Fig. 4.*

5. Conclusions and Perspectives

In this paper, a system achieving Speaker Verification based on BNs is presented. This system infers the Bayesian network structure automatically from the data. Also, it uses the independence relations obtained for integrating all the information presented on the speech signal in a single probability distribution. It shows that BNs are a flexible mathematical tool that can help to modelize information from different aspects of the speech signal. The physical interpretation given to the equations describing the structure suggests that the learning algorithms for BN are able to adequately infer the relations present in data. The perspectives for this work are important because of the flexibility of BNs. We expect further improvements from different research algorithms in the network structure learning and from the augmentation of parameters.

6. References

[1] Carey, M.J., Parris, E.S., Lloyd-Thomas, H., and Bennett, S., *Robust prosodic features for speaker identification*, International Confrence on Spoken Language Processing, vol.3, pp. 1800-1803, October 1996.

[2] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

[3] Murphy, K. P., *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph.D. Dissertation, University of California, Berkeley, Fall 2002.

[4] Kim, K.P., and Pearl, J., *A Computational Model for Combined Causal and Diagnostic Reasoning in Inference System*, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) Morgan Kaufmann Publishers, San Mateo, CA, 190:193, 1983.

[5] Cooper, G.F., *The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks*, In Artificial Intelligence, 42:393-405, 1990.

[6] MacKay, D., *Introduction to Monte Carlo Methodes*, In M. Jordan, editor, Learning in Graphical Models, MIT Press, 1998.

[7] Castillo, E., Gutiérrez, J.M., and Haidi, A.S., *Parametric Structure of Probabilities in Bayesian Networks*, Lecture Notes in Artificial Intelligence, 956:89-98, 1995.

[8] Lauritzen, S.L., and Spiegelhalter, D.J., *Local Computations with Probabilities on Graphical Structures and their Applications to Expert Systems*, Journal of the Royal Statistical Society (1988), Series B, 50:157-224.

[9] Fisher, D., and Lenz, H.J., *Learning from Data: Artificial Intelligence and Statistics V (Lecture Notes in Statistics)*, Springer Verlag (Vol 112), New York, 1996.

[10] Castillo, E., Gutiérrez, J.M., and Hadi, A.S., *Expert Systems and Probabilistic Network Models*, Springer Verlag, New York, 1997.

[11] Friedman, N., and Koller, D., *Being Bayesian about Network structure*, UAI, 2000.

[12] Cooper, G.F., and Herskovits, E., *A Bayesian Method for the Induction of Probabilistic Networks from Data*, Machine Learning, 9:309-347, 1992.

[13] Heckerman, D., *A tutorial on lerning with Bayesian Network structures*, Lerning in Graphical Models, MIT Press, 1998.

[14] Dempster, A.P., Laird, N.M., and Rubin D.B., *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society, Series B,34:1-38, 1997.

[15] *Switchboard Corpora (LDC)*, http://www.ldc.upenn.edu/

[16] Thévenaz, P., *Résidu de Prédiction Linéaire et reconnaissance de locuteurs indépendante du texte*, Ph.D. Thèse. Université de Neuchâtel, Institut de Microtechnique, 1993.

[17] Faúndez-Zanuy, M., and Rodríguez-Porcheron, D., *Speaker Recognition Using Residual Signal of Linear and Nonlinear Prediction Models*, CICYT TIC97-1001-C02-02.

[18] Murphy, K.P., *The Bayes Net Toolbox for Matlab*, Computing Science and Statistics: Proceedings of the Interface, volume 33, 2001.

[19] Bilmes, J., and Zweig, G., *The Graphical Models Toolkit: An open source software system for speech and time-series processing*, Proc. IEEE ICASP 2002.

Exploiting Nonacoustic Sensors for Speech Enhancement*

T.F. Quatieri, D. Messing, K. Brady, W.B. Campbell, J.P. Campbell, M. Brandstein, C.J. Weinstein**
J.D. Tardelli, and P.D. Gatewood***

*MIT Lincoln Laboratory***
Lexington, MA
[quatieri,dmessing,kbrady,wcampbell,jpc,msb,cjw]
@ll.mit.edu

*ARCON****
Waltham, MA
[jdt,pdg]@arcon.com]

ABSTRACT

Nonacoustic sensors such as the general electromagnetic motion sensor (GEMS), the physiological microphone (P-mic), and the electroglottograph (EGG) offer multimodal approaches to speech processing and speaker and speech recognition. These sensors provide measurements of functions of the glottal excitation and, more generally, of the vocal tract articulator movements that are relatively immune to acoustic disturbances and can supplement the acoustic speech waveform. This paper describes an approach to speech enhancement that exploits these nonacoustic sensors according to their capability in representing specific speech characteristics in different frequency bands. Frequency-domain sensor phase, as well as magnitude, is found to contribute to signal enhancement. Preliminary testing involves the time-synchronous multi-sensor DARPA Advanced Speech Encoding Pilot Speech Corpus collected in a variety of harsh acoustic noise environments. The enhancement approach is illustrated with examples that indicate its applicability as a pre-processor to low-rate vocoding and speaker authentication, and for enhanced listening from degraded speech.

1. INTRODUCTION

Linear filtering-based algorithms for additive noise reduction include spectral subtraction, Wiener filtering, and their adaptive renditions [NRC, 1989]. Nonlinear techniques have also arisen including wavelet-based noise reduction systems [Donaho and Johnson, 1994] and suppression filters based on auditory models [Hanson, 1995]. Although promising, these methods suffer from a variety of limitations such as requiring estimates of the speech spectrum and speech activity detection from a noisy acoustic waveform, distortion of transient and modulation signal components, and the lack of a phase estimation methodology.

In this paper, we present an alternative approach to noise suppression that capitalizes on recent developments in nonacoustic sensors that are relatively immune to acoustic background noise, and thus provide the potential for robust measurement of speech characteristics [Ng et al, 2000]. The effort focuses on the general electromagnetic motion sensor (GEMS) [Burnett et al, 1999], but also investigates the physiological microphone (P-mic) [Scanlon, 1998], and the electroglottograph (EGG) [Rothenberg, 1992]. These sensors can directly measure functions of the speech glottal excitation and, to a lesser extent, attributes of vocal tract articulator movements.

In Section 2 of this paper, we first formulate the enhancement problem of interest and review a specific noise reduction algorithm based on an adaptive Wiener filter [Quatieri and Dunn, 2002]. Section 3 describes the GEMS, P-mic and EGG nonacoustic sensors, as well as the DARPA Advanced Speech Encoding Pilot Speech Corpus recorded in a variety of harsh noise environments. In Section 4, we present an approach to speech activity detection based on different sensor modalities. Section 5 introduces a general multimodal methodology for improving speech spectral magnitude and phase recovery in the context of our specific adaptive suppression framework. Section 6 provides a complete multimodal speech enhancement scheme that utilizes the GEMS, P-mic, and acoustic sensors in different frequency bands. In this section, we also discuss the applicability of the proposed enhancement system to pre-processing for speech encoding and speaker authentication. Finally, in Section 7, we summarize and give future directions.

*This work is sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2. FRAMEWORK

2.1 Baseline suppression filter

Let $y[n]$ be a discrete-time noisy sequence

$$y[n] = x[n] + b[n]$$

where $x[n]$ is the desired sequence and $b[n]$ is uncorrelated background noise, both of which are assumed for the moment to be wide-sense stationary random processes with corresponding spectral density functions given by $S_x(\omega)$ and $S_b(\omega)$, respectively. One approach to recovering the desired signal is to find a linear filter $h[n]$ such that the sequence $\hat{x}[n] = y[n] * h[n]$ minimizes the expected value of $(\hat{x}[n] - x[n])^2$. The solution to this optimization problem in the frequency domain is given by

$$H(\omega) = \frac{S_x(\omega)}{S_x(\omega) + S_b(\omega)}$$

which is referred to as the Wiener filter. The required spectral densities can be estimated by averaging over multiple frames that contain only the desired signal $x[n]$ or background signal $b[n]$. Typically, however, the desired signal is nonstationary with short-duration, transient components with spectra difficult to measure, requiring an average to be essentially instantaneous.

Consider then a signal $y[n]$ processed at frame interval L samples with short-time Fourier transform $Y(kL,\omega) = X(kL,\omega) + B(kL,\omega)$ where $X(kL,\omega)$ and $B(kL,\omega)$ denote the short-time Fourier transforms of $x[n]$ and $b[n]$, respectively. And suppose we have available an estimate of the Wiener filter on frame $k-1$, denoted by $H(k-1,\omega)$. We assume that the background noise spectral density, $S_b(\omega)$, is known or estimated by averaging spectra over a given background noise region. Assuming that the desired signal $x[n]$ is nonstationary, one approach to obtain an estimate of its time-varying spectral density on the k^{th} frame uses the Wiener filter $H(k-1,\omega)$ to enhance the current frame. This operation yields an enhanced spectral estimate $\hat{X}(kL,\omega) = H(k-1,\omega)Y(kL,\omega)$ which is then used to update the Wiener filter for the next frame.

An approach to slow down a rapidly-varying $\hat{X}(kL,\omega)$, while avoid the blurring of time-varying sounds, is to temporally smooth $\hat{X}(kL,\omega)$ using a time constant that changes with the degree of stationarity of the signal [Quatieri and Baxter, 1997]. Although this filter adaptation results in relatively more noise in non-stationary regions, there is evidence that, perceptually, noise is masked by rapid spectral changes and accentuated in otherwise stationary regions [Quatieri, 2002].

A measure of the degree of stationarity is obtained through a spectral derivative defined for each frame as the mean-squared difference between two consecutive short-time spectral magnitude measurements of $y[n]$. The smooth spectral derivative is then mapped to a time-varying time constant $\tau(k)$. The use of spectral change in the Wiener filter adaptation, as well as a number of refinements to the adaptivity, including iterative re-filtering and background-noise adaptation, helps avoid blurring of temporal fine structure [Quatieri and Baxter, 1997], [Quatieri and Dunn, 2002]. We can further improve adaptivity by providing distinct Wiener filters, one during background and one during speech, thus alleviating the need to re-adapt across speech/background boundaries. The inclusion of background-noise adaptation and distinct, state-dependent Wiener filters requires that we perform speech activity detection to determine which frames in a signal contain speech and background noise and which frames contain background noise only. Finally, an enhanced speech waveform is obtained by overlap-add synthesis from the modified short-time sections. An analysis window of 12 ms and frame interval of 2 ms are used. The baseline noise-suppression algorithm is illustrated in Figure 1.

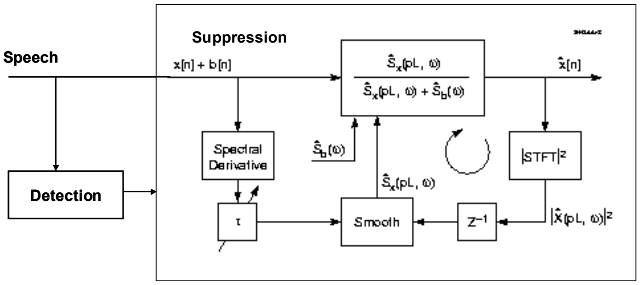


Figure 1: Noise reduction algorithm based on spectral change

2.2 Limitations

We have applied the above adaptive suppression algorithm to noise-corrupted speech under different background noise conditions, including fan, automobile, road, and cellular noise at a variety of signal-to-noise ratios (SNR) [Quatieri and Dunn, 2002]. In informal listening, the reconstructions are judged to be "crisp" corresponding to good temporal resolution of rapidly-moving and short-duration speech events. The background noise is significantly suppressed and of high quality without musicality. Nevertheless, we have observed numerous limitations particularly in environments at very low SNR, including:

Speech Activity Detection: The accuracy of speech activity detection decreases with decreasing SNR, especially with non-stationary noise. Even when correct, only one detection decision is made per frame. Ideally, multiple decisions should be made across the speech band to determine if a frequency interval is dominated primarily by speech or noise energy.

Magnitude estimation: Requirement of the speech spectrum makes it difficult to form an accurate Wiener filter, especially in low SNR frequency regions. Both the background noise and acoustic transducer can contribute to a band-dependent low-SNR.

Phase estimation: Although the Wiener filter represents a least-squared error (LSE) solution, we have found that this solution is not always perceptually "optimal". The LSE solution yields a zero-phase suppression filter so that the phase of the short-time noisy signal is left intact. In high noise conditions, phase noise is frequency-dependent and audible.

In this paper, we address these limitations in the use of measurements from nonacoustic sensors.

3.0 NONACOUSTIC SENSORS AND MEASUREMENTS

3.1 GEMS

The general electromagnetic motion sensor (GEMS) measures tissue movement during voiced speech, i.e., speech involving vocal chord vibrations [Burnett et al, 1999]. An antenna is typically strapped or taped on the throat at the laryngeral notch, but also can be attached at other facial locations. This sensor emits an electromagnetic signal that penetrates the skin and reflects off the speech production anatomy such as the tracheal wall, the vocal folds, or the vocal tract wall. Because signals collected from a GEMS device depend on the tissue movement in the speech production anatomy, it is relatively immune to degradation from external acoustic noise sources.

During voiced speech, GEMS records quasi-periodic electromagnetic signals due to vibration of the speech production anatomy. When placed at the larynx, quasi-periodic measurements are found during vowels, nasals, and voiced consonants including prior to and following the burst in voiced plosives, i.e., during voice bars. Single pulses have also been observed sporadically from the GEMS measurement at the burst in unvoiced plosive consonants.

3.2 P-mic

The physiological microphone (P-mic) sensor is composed of a gel-filled chamber and a piezoelectric sensor behind the chamber [Scanlon, 1998]. Vibrations that permeate the liquid-filled chamber are measured by the piezoelectric sensor that provides an output signal in response to applied forces that are generated by movement, converting vibrations traveling through the liquid-filled chamber into electrical signals. The liquid filled chamber is designed to have poor coupling between ambient background noise and the fluid-filled pad thus attenuating vibrations of unwanted ambient background noise.

Like the GEMS sensor, the P-mic can be strapped or taped on various facial locations. The P-mic at the throat measures primarily vocal fold vibrations with quasi-periodic measurements similar to that of GEMS. The P-mic signal at the throat, however, contains some low-pass vocal tract formants with bandwidths wider than normal. Other facial locations can provide additional vocal tract characteristics. The P-mic located on the forehead, for example, gives significant vocal tract information but is far less noise-immune than the P-mic at the throat in severe environments.

3.3 EGG

The electroglottograph (EGG) [Rothenberg, 1992] sensor measures vocal fold vibrations by providing an electrical potential (of about one volt rms and two-to-three megahertz) across the throat at the level of the larynx. With a pair of gold-plated electrodes, the sensor measures the change of impedance over time. When the vocal folds are closed, the impedance is decreased; when they are open, the impedance is increased. Thus, the opening and closing of the vocal folds, present in voiced speech, are measured by the EGG.

3.4 Corpus collection

An extensive multi-sensor speech corpus was collected from ten male and ten female talkers. Scripted phonetic, word and sentence material along with conversational material were generated by each talker. These materials were generated in nine different acoustic noise environments. The corpus was collected in two sessions (on two different days). Speakers were exposed to a variety of noise environments including both benign and severe cases. Six of the environments represented three acoustic environments with each presented at two intensity states. The presentation levels for these states differed by 40 dB SPL. Specific environments are quiet, office (56 dB), MCE (mobile command enclosure, 79 dB), M2 Bradley Fighting Vehicle (74 dB and 114 dB), MOUT (military operations in urban terrain, 73 dB and 113 dB), and a Blackhawk helicopter (70 dB and 110 dB). We call these environments (with L indicating low noise and H indicating high noise) quiet, office, MCE, M2L, M2H, MOUTL, MOUTH, BHL and BHH, respectively.

For each talker and environment, combination time-synchronous data was collected from up to seven separate sensors. These sensors consisted of the previously introduced GEMS, P-mic and EGG. Data was also collected from two acoustic microphones, a high quality B&K calibration microphone and an environment specific "resident" microphone. The resident microphone was typically the first-order gradient noise-cancellation microphone used for normal communications in that specific environment.

One GEMS and one EGG were located near the talker's larynx. Careful attention was given to tuning the GEMS sensor and in optimizing its placement. The GEMS was considered the primary sensor during the corpus collection. A specific talker's neck and shoulder geometry often required that tradeoffs be made in the placement of the secondary sensors in order to optimize the GEMS signal. Two P-mics were used, one located in the vicinity of the talker's larynx and the other on the talker's forehead.

Due to the acoustic presentation levels of some of the noise environments, all talkers used the acoustic protection systems typical of each specific noise environment. This normally consisted of some type of communication headset that provided noise attenuation on the order of 20 dB. Human subject procedures were followed carefully and noise exposure was monitored.

The complete corpus consists of up to eight channels of data from approximately twenty minutes of speech material in each of nine acoustic noise environments from each of the twenty talkers. All sensor data was sampled at 48 kHz, though the nonacoustic data was downsampled to 16 kHz for space considerations. The full corpus takes approximately 70 GB of storage.

4.0 SPEECH ACTIVITY DETECTION

Speech activity detection is used to identify which segments of an acoustic waveform contain speech with background noise and which contain only background noise. This detection is useful because it allows state-dependent processing, as is performed in the adaptive suppression filter of Section 2. As also noted in Section 2, however, the accuracy of detectors based on the acoustic waveform decreases with decreasing SNR and, even when correct, only one detection decision is made per frame, thus not accounting for a frequency-dependent SNR.

4.1 Multi-sensor detection

Our approach to circumventing the speech detection problem caused by noise in acoustic waveforms is to use the waveforms from other sensors that are less sensitive to acoustic background noise. The GEMS and EGG sensors, for example, are robust at detecting voiced speech, both during vowels and voicing associated with voiced consonants. Although these sensors are poor in measuring the noise component of unvoiced speech sounds, relative to their acoustic counterpart, they give more accurate speech activity detection resulting in increased segmental signal-to-noise ratio in harsh environments [Messing, 2003]. The P-mic sensor is less accurate at detecting voiced speech since it is not entirely immune to acoustic noise. Under certain noise conditions and placements, on the other hand, it can detect the noise component of unvoiced speech.

It follows that one approach to improve detection, relative to that from the acoustic signal, is to perform voiced speech detection using the GEMS or EGG sensor and then, given this voiced speech detection decision, use the P-mic sensor waveform to decide on the presence of unvoiced speech. This fusion has been found to improve speech activity detection of both voiced and unvoiced speech relative to using any one sensor alone [Messing, 2003]. An example of this fusion-based detection is illustrated in Figure 2 where the GEMS and P-mic sensors are used.

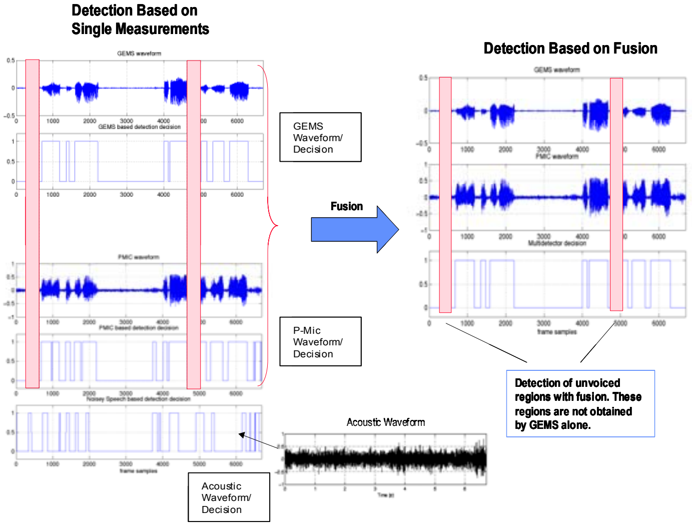


Figure 2: Illustration of multi-detector fusion using the GEMS and P-mic sensors. In this case, the sensor signals are from the M2H environment. For test purposes, the acoustic signal being enhanced is from the acoustic B&K mic and "truth" is assumed as the output of the corresponding noise-cancelling resident mic.

Although this style of detection can outperform acoustic-based detection in noise, it does not address the limitation of a binary detection decision per frame. In the next section, we propose a detector that provides frequency-dependent speech activity evaluation.

4.2 Multi-class detection

There are numerous speech events that may be characterized by the pattern of signal energy across specific frequency bands. For example, unvoiced fricative and plosive sounds are high in frequency, while nasals and the voice-bar components of voiced plosives are low in frequency. For these cases a single speech activity decision limits the performance of the adaptive suppression filter of Section 2. Consequently, we refine our multi-sensor detector by exploiting the propensity of the various sensors to detect speech events in different frequency bands. The resulting scheme detects three speech classes: (1) Voiced = Speech present in low- and high-frequency bands; (2) Low-voiced (including nasals and voice bars) = Speech is present in low-frequency band only; and (3) Unvoiced = Speech is present in high-frequency band only. A background state is declared when speech is not present in either band. The four-class detection scheme is illustrated in Table I. According to the motivation given below, the low band is selected as [0, 500] Hz and the high band as [3, 5] kHz.

	Contains High Frequency Speech Content?	Contains Low Frequency Speech Content?
Low-Voiced	No	Yes
Unvoiced	Yes	No
Voiced	Yes	Yes
Background	No	No

Table I: Four speech-class detection scheme.

The multi-class decisions are based on a detection scheme much like the above fused GEMS and P-mic detectors. Rather than using the P-mic signal, which does not robustly provide high-frequency signal estimates in harsh conditions, we use the wideband signal from the resident-mic in the ASE Pilot Speech Corpus. For certain harsh conditions of interest, the SNR is large for frequencies within about a 3 kHz to 8 kHz range. This is advantageous because several consonants such as an /s/, /sh/, /ch/, and /th/ contain significant energy in this frequency region. In practice, both voicing and unvoicing above 3 kHz were found to be detectable. On the other hand, for these same conditions, the resident-mic has a poor SNR for low frequencies, particularly below about 500 Hz and thus is not reliable at detecting low-frequency events such as nasals and voice bars. An example is shown in Figure 3 where the GEMS signal clearly gives the presence of the low-frequency nasal /n/ and voice bar in the voiced plosive /d/ in the word "dint". The resident-mic, while not revealing the nasal and voice bar, more clearly shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/.

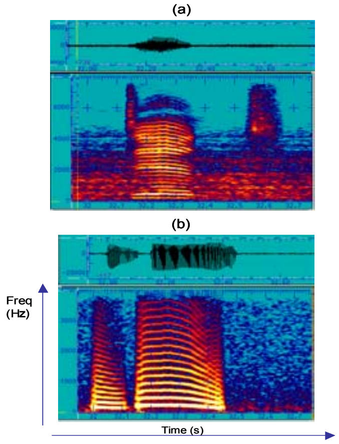


Figure 3: Waveforms (from the M2H environment) and spectrograms of the (a) resident-mic signal and (b) GEMS signal for the word "dint". The GEMS signal shows the presence of the nasal /n/ and voice bar in the initial voiced plosive /d/, while the resident-mic shows the high-frequency burst energy in the /d/ and in the unvoiced plosive /t/.

One multi-class detector that we are investigating involves first the use of GEMS, a low-pass signal by its nature, to detect voicing in low frequencies. The signal from the resident mic above 3 kHz is then used 100 ms on either side of the region detected by the GEMS-based detector, similar to the P-mic and GEMS fused-detector decision logic.

5.0 MAGNITUDE AND PHASE ESTIMATION

The nonacoustic sensors may also be exploited to aid in the suppression component of the enhancement algorithm of Section 2. In this section, we first investigate the theoretical limits of the use of an "ideal" speech magnitude and phase as components in the Wiener filter. We then look at numerous strategies to approach these limits with nonacoustic sensor measurements.

5.1 Theoretical limits

In the frequency domain, we can view the noisy speech signal in terms of its short-time Fourier transform which we write in polar form as

$$\begin{aligned} Y(k,\omega) &= X(k,\omega) + B(k,\omega) \\ &= [|X(k,\omega)| + M_e(k,\omega)]e^{j[\theta_x(k,\omega)+\theta_e(k,\omega)]} \end{aligned}$$

where $M_e(k,\omega)$ and $\theta_e(k,\omega)$ are short-time magnitude and phase noise terms, respectively. One approach to viewing the "theoretical limit" in magnitude and phase estimation is to make each noise term zero. We have performed these replacements over a range of SNRs and have made the following observations:

Ideal magnitude: When replacing the noisy magnitude with its ideal form, phase noise persists aurally and this noise increases with decreasing SNR. A sub-optimal performance bound can be achieved by constructing a Wiener filter using the ideal speech magnitude. The resulting enhanced signal lies aurally between that using the ideal magnitude and noisy magnitude.

Ideal Phase: When replacing the noisy phase with its ideal form, an increasing noise reduction occurs aurally with decreasing SNR.

Although we have not yet made quantitative measurements or performed formal listening tests, our anecdotal results are consistent with those of Vary who found the perceptual importance of phase with respect to signal quality to increase with decreasing SNR, in particular for SNR < 3 dB [Vary, 1985]. We have also performed analogous experiments to the above with the magnitude estimated from the adaptive suppression filter of Section 2. Using our estimated magnitude in place of the noisy magnitude yields a reduction in noise residual but less than with the ideal magnitude, as expected. Likewise, we find that when replacing the noisy phase with its ideal form, an increasing noise reduction occurs aurally with decreasing SNR.

5.2 Strategies for exploiting nonacoustic sensors

The previous section indicates that we have not reached performance bounds with magnitude estimation and that we can gain considerably with phase estimation under harsh background noise conditions. In this section, we describe a general approach to capitalize on these observations using nonacoustic sensors.

Magnitude estimation: We have seen that a drawback of Wiener filtering is the need to estimate time-varying speech spectra from the noisy acoustic waveform. An alternative strategy is to estimate short-time speech spectra for the Wiener filter from a nonacoustic-sensor signal and the background noise spectrum from the acoustic signal. Since the estimate of the speech spectrum depends on the particular nonacoustic sensor and its placement, and its fidelity can be band-dependent, our general strategy is to construct a speech spectral estimate from a fusion of components of nonacoustic and acoustic signals. Alternatively, in contrast to constructing the ideal Wiener filter, one can aim for the ideal speech spectrum by replacing bands of the acoustic signal with those of a nonacoustic signal where appropriate. Since nonacoustic signals, although relatively noise-immune, can themselves be degraded, we apply noise suppression to these signals prior to fusion. An example of this strategy is given in Figure 4 which compares the baseline suppression on a resident-mic acoustic signal with a fusion of the enhanced acoustic signal above 500 Hz with the enhanced P-mic signal below 500 Hz. In this example word "zed", voicing during the voiced fricative /z/ and the voice bar for voiced plosive /d/ have been approximately restored by the low-band P-mic signal.

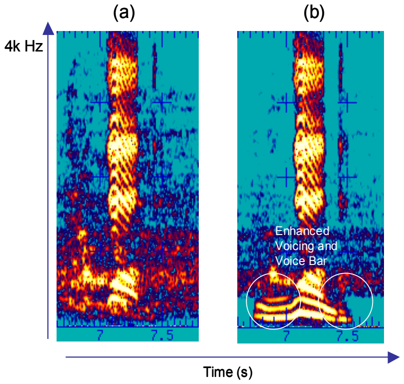


Figure 4: Spectrogram comparison of (a) enhanced resident-mic acoustic signal (from the M2H environment) and (b) fusion of the enhanced acoustic signal above 500 Hz with the enhanced P-mic nonacoustic signal below 500 Hz. In this example word "zed", voicing during the voiced fricative /z/ and the voice bar for voiced plosive /d/ have been approximately restored by the low-band P-mic signal.

Phase estimation: In phase estimation, our goal is to use nonacoustic sensors to recover the short-time Fourier transform phase of the speech waveform. Based on a linear speech production model, the speech phase consists of the sum of the excitation and vocal tract components. Generally, we have found that different sensors contain different speech phase components. The GEMS (placed at the larynx in the ASE Pilot Speech Corpus) and EGG contain primarily excitation phase during voicing. The P-mic contains excitation phase, but also vocal tract phase over selected time-frequency regions, depending on its location. Figure 5 gives an example of the use of phase from the P-mic signal in place of that of the noisy phase from the acoustic signal, enhanced by the baseline suppression algorithm. In this case, the P-mic was placed a fair distance above the larynx, giving significant vocal tract phase, as well as excitation phase. Replacing the noisy phase of the Wiener-enhanced speech by the phase of the P-mic signal over the full 4 kHz band results in additional noise reduction and cleaner harmonic structure than from the baseline suppression filtering.

Observe from Figure 5 an apparent paradox: *Modifying the short-time phase modifies the short-time magnitude.* This paradox is resolved by recalling that the enhanced signal is synthesized by an overlap-add scheme. Therefore, cleaning the phase of the noisy short-time Fourier transform may reduce noise in the magnitude of the short-time Fourier transform of the enhanced synthesized signal by virtue of overlapping and adding. In addition, the spectrogram views the synthesized

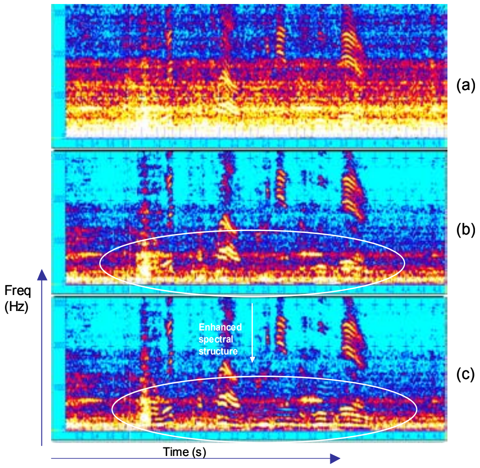
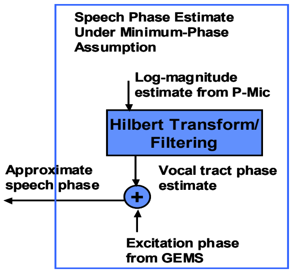


Figure 5: Spectrograms illustrating use of the P-mic phase in noise suppression: (a) Original noisy resident-mic signal (from M2H environment); (b) (a) enhanced by baseline suppression filter; (c) Use of P-mic phase in place of noisy resident-mic phase of (b). Acoustic noise in this case rolls off at about 2000 Hz.

waveform through a window and frame (in this example, 25 ms and 10 ms, respectively) different from that of the enhancement analysis/synthesis (12 ms and 2 ms, respectively).

There are many different scenarios of this phase recovery scheme. When reliable vocal-tract phase is not available, one can use the excitation phase solely from a nonacoustic sensor. Although this phase replacement can enhance the speech signal in the sense of giving the perception of less noise, it often leads to an annoying "buzzy" quality due to excessive phase coherence. A preferred strategy is to use the excitation phase from the GEMS or EGG and a vocal tract phase estimate from the resident-mic or P-mic derived under a minimum-phase assumption [Quatieri, 2002]. As illustrated in Figure 6 with the P-mic sensor, the synthetic vocal phase is constructed through a Hilbert transform and filtering to remove an excitation contribution, implemented by homomorphic-filtering of the nonacoustic signal. This allows the possibility of an enhanced vocal tract phase from an enhanced P-mic spectral magnitude, falling between the ideal phase and the original noisy phase. More generally, this constructed phase can be fused over selected bands with the phase of other sensors.

Figure 6: Phase construction by addition of the GEMS phase and a synthetic phase derived from the P-mic signal under a minimum-phase assumption. The Hilbert transform/filtering module implements homomorphic filtering of the enhanced P-mic signal


Speech Phase Estimate Under Minimum-Phase Assumption
Log-magnitude estimate from P-Mic
Hilbert Transform/ Filtering
Approximate speech phase
Vocal tract phase estimate
+
Excitation phase from GEMS

6.0 Composite System Example and Implications

We are investigating a number of system configurations based on the speech activity detection and magnitude and phase estimation schemes of Sections 4 and 5. One particular configuration is shown in Figure 7. In this scheme, we use our multi-class detector, based on the GEMS and the wide-band resident mic. The suppression uses the P-mic signal for the low-band [0, 500] Hz and the resident-mic signal above 500 Hz. The GEMS and P-mic are used for a synthetic phase over the range [500, 1200] Hz. As seen in Figure 7, we have introduced a spectral sharpening module which narrows formant bandwidths and adds a pre-emphasis, accounting for formant widening by the P-mic due to energy loss through the skin and accounting for there being no acoustic radiation as in the acoustic waveform that emanates through free space from the lips.

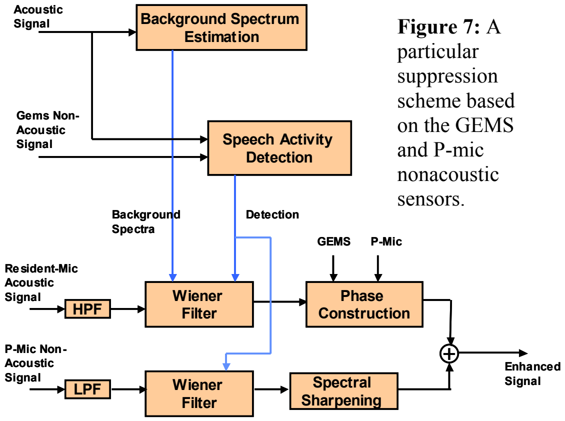


Figure 7: A particular suppression scheme based on the GEMS and P-mic nonacoustic sensors.

An example using this configuration is illustrated in Figure 8 in two steps: (1) Combined Wiener filtering of the high-pass resident-mic and the low-pass P-mic signals, and (2) Inclusion of the phase construction of Figure 6.

Observe that the introduction of the enhanced low-passed P-mic signal has provided significant voice-bar and nasal consonant components lost by the resident-mic. The introduction of the synthetic phase derived from the GEMS and P-mic has improved the visual clarity of the harmonic and formant structure in the mid-frequency range, relative to the noisy phase of the original acoustic signal. Informal listening to a variety of passages, under the M2H condition, indicates improved quality, and potentially improved intelligibility, for enhanced listening. In addition, we cite two other application areas:

Speaker authentication: The nonacoustic sensors appear to provide an accentuation of low-frequency events such as voice bars and nasals. We have observed a strong speaker-dependence of the duration, strength, and spectral character of these events in nonacoustic signals, and thus our enhanced, fused signals may provide information for speaker authentication [Campbell et al, 2003] not represented with the baseline suppression.

Speech encoding: Including the synthetic phase of Figure 6 appears as a small effect in some cases, especially in bands of high- to moderate SNR. In the encoding application, however, this effect multiplies. Figure 9 shows the result of MELP encoding [McCree et al, 1996] the signal of Figure 8, before and after including the synthetic phase. The introduction of phase has significantly improved harmonic structure of the signal not only in the [500, 1200] Hz range where the phase is replaced, but also through the entire speech band, and with informal listening this corresponds aurally to a perceived improved "clarity" and lessening of coding artifacts.

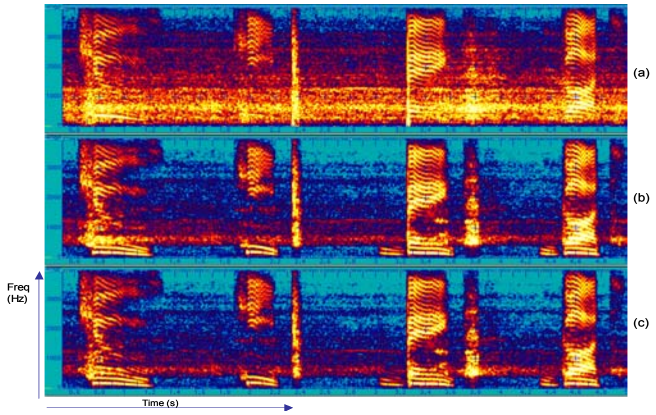


Figure 8: Example spectrograms of enhancement using the suppression scheme of Figure 7: (a) Original noisy resident-mic; (b) Inclusion of low-passed P-mic; (c) Inclusion of synthetic phase. The utterance consists of the word stream "choose-keep-bank-got" in the M2H environment.

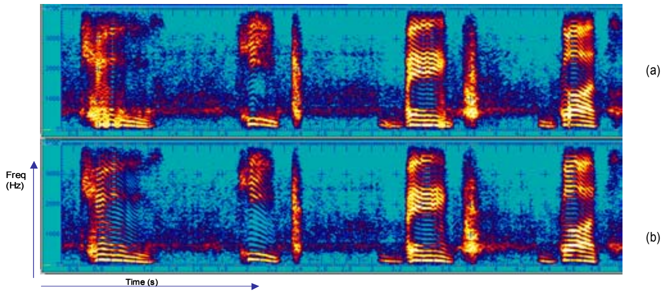


Figure 9: Example spectrograms of enhancement using the suppression scheme of Figure 7 after MELP encoding of signals in Figure 8: (a) Original noisy phase; (b) Introduction of synthetic speech phase in band [500, 1200] Hz.

7.0 SUMMARY AND DISCUSSION

In this paper, we presented an approach to noise suppression that capitalizes on recent developments in nonacoustic sensors that are relatively immune to acoustic background noise. The GEMS, P-mic, and EGG nonacoustic sensors were considered. These sensors can directly measure the speech glottal excitation but also directly measure speech attributes such as vocal tract articulators. The sensors were exploited to improve a particular noise suppression system based on a Wiener filter that adapts to spectral changes in the speech and background noise signals. Different aspects of acoustic and nonacoustic signals were used according to their capability in representing specific speech characteristics. Frequency-domain sensor phase, as well as magnitude, was found to contribute to improved signal enhancement within different frequency bands. Preliminary testing involved the time-synchronous multi-sensor DARPA ASE Pilot Speech Corpus collected in harsh acoustic noise environments. The enhancement approach was illustrated with examples and shown to have potential as a pre-processor to low-rate vocoding and speaker authentication, as well as for enhanced listening.

Next steps include perceptual testing of enhanced signals, both uncoded and MELP-encoded, using a formal diagnostic rhyme test (DRT), and further development of different configurations based on nonacoustic sensor placements. We are also in the process of applying the enhancement algorithms of this paper as pre-processing for speaker authentication with the DARPA ASE Pilot Speech Corpus [Campbell et al, 2003].

References

[Burnett et al, 1999] G.C. Burnett, J.F. Holzrichter, T.J. Gable, and L.C. Ng, "The Use of Glottal Electromagnetic Micropower Sensors (GEMS) in Determining a Voiced Excitation Function," presented at the 138th Meeting of the Acoustical Society of America, November 2, 1999, Columbus, Ohio.

[Campbell et al, 2003] W.M. Campbell, T.F. Quatieri, J.P. Campbell, and C.J. Weinstein, "Multimodal speaker authentication using nonacoustic sensors," Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.

[Donahue and Johnson, 1994] D. Donahue and I. Johnson, "Ideal denoising in an orthonormal basis chosen from a library of bases," C.R. Academy of Science}, Paris, vol. 1, no. 319, pp. 1317-1322, 1994.

[Hansen and Nandkumar, 1995] J.H. Hansen and S. Nandkumar, "Robust Estimation of Speech in Noisy Backgrounds Based on Aspects of the Auditory Process," JASA, Vol. 97, No. 6, June 1995.

[McCree et al, 1996] A. McCree, K. Truong, E.B. George, T.P. Barnwell, and V, Viswanathan, "A 2.4 kbit/s MELP coder candidate for the new US Federal standard," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Atlanta, GA, vol. 1, pp. 200-203, May 1996.

[Messing, 2003] D. Messsing, Noise suppression using spectral magnitude phase from non-air-acoustic sensors, MS Thesis, MIT, August 2003.

[Ng et al, 2000] L.C. Ng, G.C. Burnett, J.F. Holzrichter, and T.J. Gable, "Denoising of human speech using combined acoustic and EM sensor signal processing," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000.

[NRC, 1989] "Removal of noise from noise-degraded speech," Panel on removal of noise from speech signal, National Academy Press, Washington, D.C, 1989.

[Quatieri and Baxter, 1997] T.F. Quatieri and R.A. Baxter, "Noise reduction based on spectral change," *IEEE 1997 Workshop on Appl. of Signal Processing to Audio and Acoustics*, pp. 8.2.1-8.2.4, New Paltz, NY, October 1997.

[Quatieri and Dunn, 2002] T.F. Quatieri and R.B. Dunn, "Noise reduction based on auditory spectral change," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* Orlando, FL, May 2002.

[Quatieri, 2002] T.F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 2002.

[Rothenberg, 1992] M. Rothenberg, "A multichannel electroglottograph," *J. of Voice*, vol. 6, no. 1, pp. 36-43, 1992.

[Scanlon, 1998] M.V. Scanlon, "Acoustic sensor for health status monitoring," *Proceedings of IRIS Acoustic and Seismic Sensing*, vol. 2, pp. 205-222, 1998.

[Vary, 1985] P. Vary, "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits," *Signal Processing*, vol. 8, pp. 387-400, 1985.

Segmental Score Fusion for Text-independent Speaker Verification

Asmaa El Hannani*§, Dijana Petrovska-Delacrétaz*, Raphaël Blouet† and Gérard Chollet†
*DIVA group, University of Fribourg, Fribourg, Switzerland;
Email: {asmaa.elhannani,dijana.petrovska}@unifr.ch
§Supported by the Swiss National Fund for Scientific Research.
† École Nationale Supérieure des Télécommunication, dépt. TSI, 46 rue Barrault, 75634 Paris;
Email: {blouet,chollet}@tsi.enst.fr

***Abstract*— Current state-of-the-art speaker verification algorithms use Gaussian Mixture Models (GMM) to estimate the probability density function of the acoustic feature vectors. Previous studies have shown that phonemes have different discriminant power for the speaker verification task. In order to better exploit these differences, it seems reasonable to segment the speech in distinct speech classes and carry out the speaker modeling for each class separately. Because transcribing databases is a tedious task, we prefer to use data-driven segmentation methods. In our previous work, we have focused on the tuning of the ALSIP data-driven segmentation method. The novelty of the proposed method is the combination of the DTW distortion measure with data-driven segmentation tools, and the use of a Logistic Regression Function to determine the optimal fusion weights of the speech segments. The performance of the proposed system is evaluated on subsets build from NIST'2001 and 2002 Evaluation data. Our results show that applying score fusion with the weights found by the Logistic Regression function leads to a better results, as compared to a simple summation of the segmental scores. Our method could also be applied to automatically remove the less significant segments (usually corresponding to "silence" segments).**

I. Introduction

Current best performing text-independent speaker verification systems are based on Gaussian Mixture Models (GMM) [1]. Speech is composed of different sounds and speakers differ in their pronunciation of these sounds. GMM could be interpreted as a "soft" representation of the various acoustic classes that make up the speakers' sounds. They do not take into account the temporal ordering of the feature vectors. The speaker verification approach described in this work is based on speech recognition, grounded on data-driven techniques that require neither phonetic nor orthographic transcriptions of the speech data. The main advantages of introducing a speech recognition stage in speaker verification experiments are: to exploit the different speaker discriminant power of speech sounds [2,3,4,5], and to benefit of some higher-level informations resulting from the segmentation.

The majority of current speech processing systems use phones (or related units) as an atomic representation of speech. Using phonetic speech units lead to efficient representation and implementation for a lot of speech processing systems. The major problem that arises when phone based systems are being developed is the possible mismatch with the data being used and the lack of transcribed databases (because transcribing speech data is an error-prone and expensive task). The set of speech units can also be learned from examples, like in data-driven approaches. In [4-6] we have proposed a new architecture for speech processing based on units acquired during a data-driven segmentation, that is not grounded on transcribed databases. These units are denoted as Automatic Language Independent Speech Processing (ALISP) units.

In [4], [5] we have already used the ALISP data-driven speech segmentation method for speaker verification. The number of classes, (8), was chosen in order to have enough data for each class, when dealing with 2 min of enrollment speech data used to build the speaker models. In those experiments, we studied speaker modeling algorithms, such as Multiple Layer Perceptrons (MLP) and GMM's. Classifying speech in only 8 speech classes, did not led to a good coherence of the speech classes.

In [17], we have used a finer segmentation of the speech data into 64 speech classes, and a Dynamic Time Warping (DTW) distortion measure for the speaker verification step. It is obvious that when using so many classes, the classical speaker modeling methods have to be redefined. Speaker modeling with GMM's is still possible, but more difficult because of the lack of client speaker data. MLP could not be applied, because of lack of sufficient data for the client class. Therefore, we have decided to use the well known Dynamic Time Warping method to evaluate the distance between two speech patterns. This method could be used independently on short and long speech data. If the two speech patterns belong to the same speech class, we could expect that the DTW distortion measure can capture the speaker specific characteristics. DTW distance measures have already been used for text-dependent speaker recognition experiments [11], [7], [12], [13]. The novelty of the proposed method is its combination with the ALISP units. In our previous work, we have focused on the tuning of the ALISP based segmentation. In [17], we have analyzed the scores used for the speaker verification on a global level, and all the speech classes were treated with the same weight. In this paper, we use the Logistic Regression [14], [15], [16] to exploit the different discriminant power of the ALISP speech classes.

The outline of this paper is the following: In Sect. II

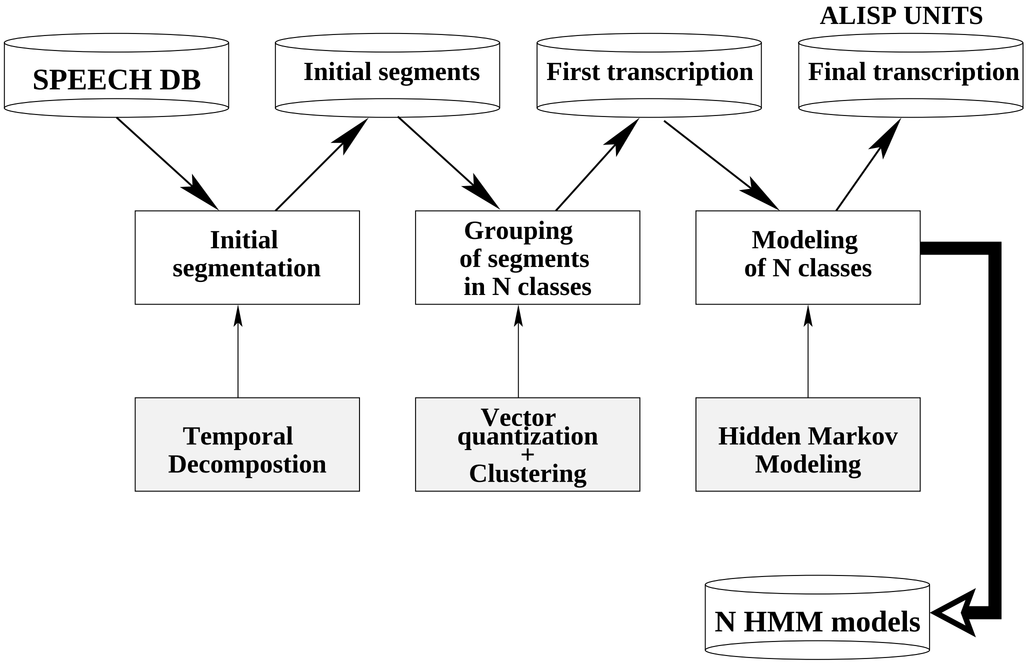


Fig. 1. Unsupervised Automatic Language Independent Speech Processing (ALISP) unit acquisition, and their HMM modeling

presents in a more detailed way the proposed method. Sect. III describes the database used and the experimental protocol. The evaluation results are reported in Sect. IV. The conclusions and perspectives are given in Sect. V.

II. Description of the Proposed System

A. Data-Driven Speech Segmentation

The steps needed to acquire and model the set of data-driven speech units, denoted here as *Automatic Language Independent Speech Processing (ALISP)* units [6], are shortly described in the following section. Instead of the widely used phonetic labels, data-driven labels automatically determined from the training corpus are used. The set of symbolic units is automatically acquired through temporal decomposition, vector quantization, segment labeling and Hidden Markov Modeling, as shown in Figure 1. After a classical pre-processing step leading to acoustic-feature vectors, temporal decomposition [8] is used for the initial segmentation of the speech data into quasi-stationary segments. At this point, the speech is segmented in spectrally stable portions. For each segment, its gravity center frame is determined. A vector quantization algorithm is used to cluster the center of gravity frames of the spectrally stable speech segments. The codebook size defines the number of ALISP symbols. The initial labeling of the entire speech segments is achieved using minimization of the cumulated distances of all the vectors from the speech segment to the nearest centroid of the codebook. The result of this step is an initial segmentation and labeling. These labels are used as the initial transcriptions of the ALISP speech units. Hidden Markov Modeling is further applied for a better coherence of the initial ALISP units. Since correct transcriptions of the evaluation data are not available, we cannot compare the correspondence of ALISP units and the usual phonetic units. We studied this correspondence for some speaker of the development set and we found that there is some evidence of correlation of phonemes and ALISP units (see Figure 2).

B. Principle of the Segmental Speaker Verification

The proposed speaker verification system is a combination of Dynamic Time Warping (DTW) distortion measure with data-driven speech segmentation based on ALISP tools. The number of speech classes used is 64, and is comparable to a pseudo-phonetic segmentation.

For the speaker verification step we use two dictionaries: the *Client-Dictionary*, composed of the segments found in the enrollment client speech data and the *World-Dictionary*, build with segments found in the speech data representing the world speakers. These dictionaries are defined during the training (also known as enrollment) phase.

During the test phase, each test speech data, Y is first segmented with the N ALISP HMM models. If we denote by M the number of total ALISP segments in the test segment Y. Y is the concatenation of M segments y_t, $t = 1, ..., M$.

In the next step of the testing phase, each of the test speech segments y_t is compared with a DTW distance measure, to the *Client-Dictionary* and to the *World-Dictionary*. This comparison is done on a per class level. For sake of simplicity, we will omit the indexes indicating the ALISP classes. The score $s(y_t)$, for each segment, is calculated as follows (see also Figure 3):

$$s(y_t) = \frac{D(y_t, y_c) - \mu_W}{\sigma_W} \qquad (1)$$

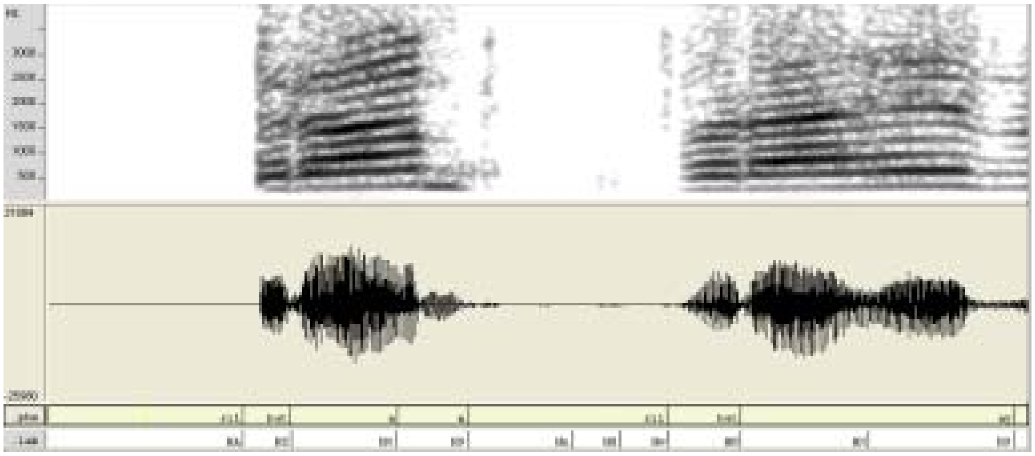


Fig. 2. Example of the ALSIP segmentation, of an excerpt of a test speech data, spoken by a male speaker.

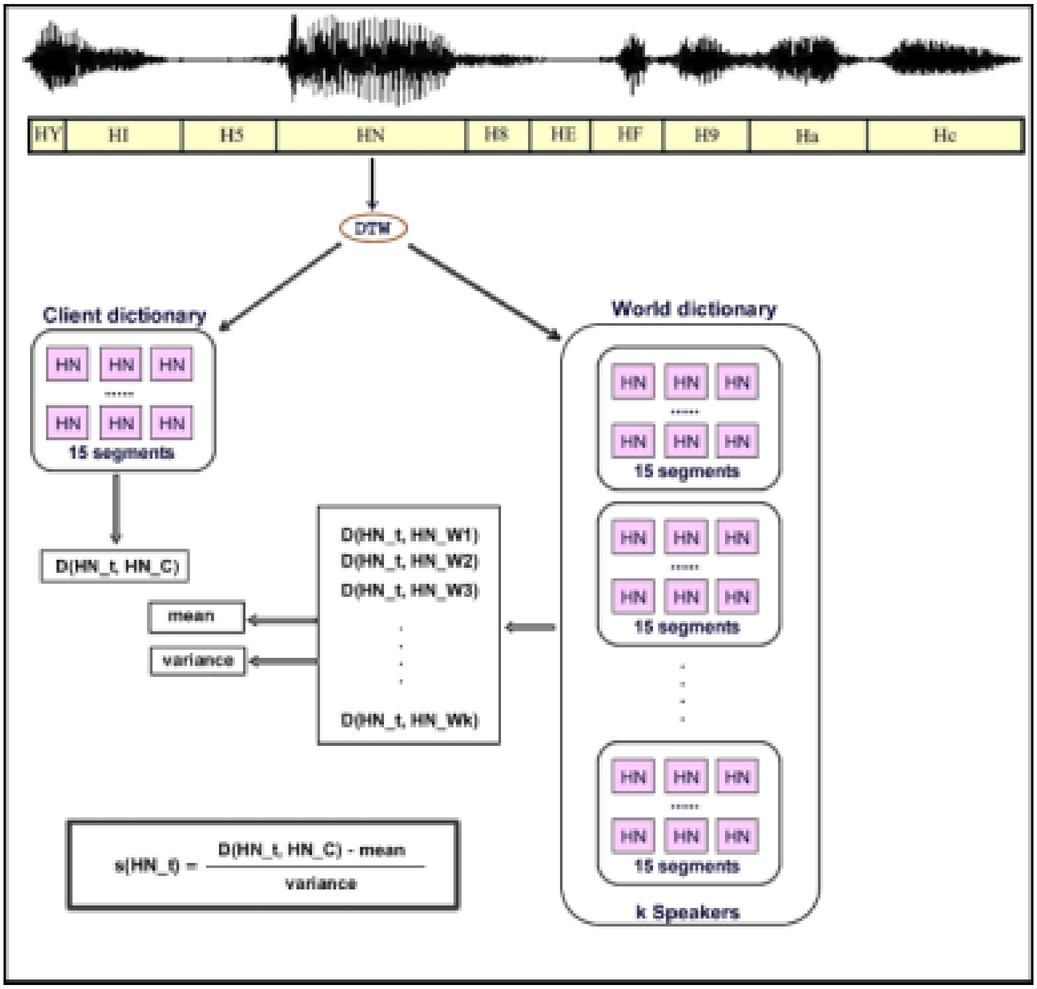


Fig. 3. Illustration of the proposed segmental method based on searching in a client and world speech dictionaries.

where $D(y_t, y_r)$ is the distance of y_t to y_r the most similar segment from the corresponding class dependent Client-Dictionary ; μ_W and σ_W are respectively denoting the mean and variance of the most similar segments from each of the W_k world speakers dictionary.

Let M be the number of segments in the test speech data. The global score of the claimed speaker is then calculated as a simple summation of the segmental scores (defined by Equation 1), and normalized by the total number of segments M:

$$S = \frac{1}{M}\sum_{t=1}^{M} s(y_t) \tag{2}$$

C. Applying Logistic Regression for Segmental Score Fusion

The system described in the previous section uses the same weight for all the segments. To exploit the different discriminant power of the speech classes, the Logistic Regression is applied. The Logistic Regression function ([15], [14], [16]) is defined as follows:

$$g(s) = \omega_0 + \sum_{t=1}^{M} \omega_t s(y_t), \tag{3}$$

$$\omega_0 = \sum_{t=1}^{M} \frac{(\mu_t^C)^2 - (\mu_t^I)^2}{\sigma_t^2} + ln\frac{P(C)}{P(I)}, \tag{4}$$

$$\omega_t = \frac{\mu_t^C - \mu_t^I}{\sigma_t^2}. \tag{5}$$

where μ_t^C and μ_t^I represent the mean of the client and impostor classes, respectively, and σ_t^2 represents their common variance. ω_t is the weight given to the segment y_t of the class t. This assumes gaussianity of the score distributions with equal variances. The global score S, can be rewritten as follows.

$$S = \frac{1}{M}\sum_{t=1}^{M} \omega_t s(y_t) \tag{6}$$

This weights are estimated from a development set.

III. Experimental Setup

The experiments described in this paper are carried out using the NIST 2001 and 2002 Evaluation data [18]. In order to evaluate the proposed method, 3 disjoint sets denoted here as: *World-ALISP-set*, *Development-set* and *Evaluation-set* are designated among the available data.

The *World-ALISP-set* (comprising data from 58 female and 57 male speakers), is a subset of the NIST'2001 data. It is used for two purposes: to build the gender dependent ALISP recognizers and to represent the world speakers. The NIST'2002 data is split into two sets: the *Development-set* (80 female and 60 male speakers), used to estimate with the Logistic Regression function, the class dependent weighing values and the *Evaluation-set* (111 female and 79 male speakers), used to test the performance of the proposed system.

The speech parameterization for the temporal decomposition is done with Linear Prediction Cepstral Coefficients (LPCC), calculated on 16 ms windows, with a 8 ms shift (this choice is due to implementation purposes). For the speech recognition with the Hidden Markov Models and for the DTW speaker verification step, we have used the Mel Frequency Cepstral Coefficients (MFCC). They are generally used for common speech and speaker verification purposes. The window and shift values are kept the same as for the LPCC parameterization.

In order to accelerate the search, we have restricted the number of speech units in the client and world dictionaries. The 15 longest segments per class and per world speakers are chosen for the *World dictionary*, and for the *Client dictionary*. During the testing phase (see Figure 3), each of the test speech segments is compared with a DTW distance measure, to the *Client Dictionary* and to the *World Dictionary*.

IV. Fusion Results

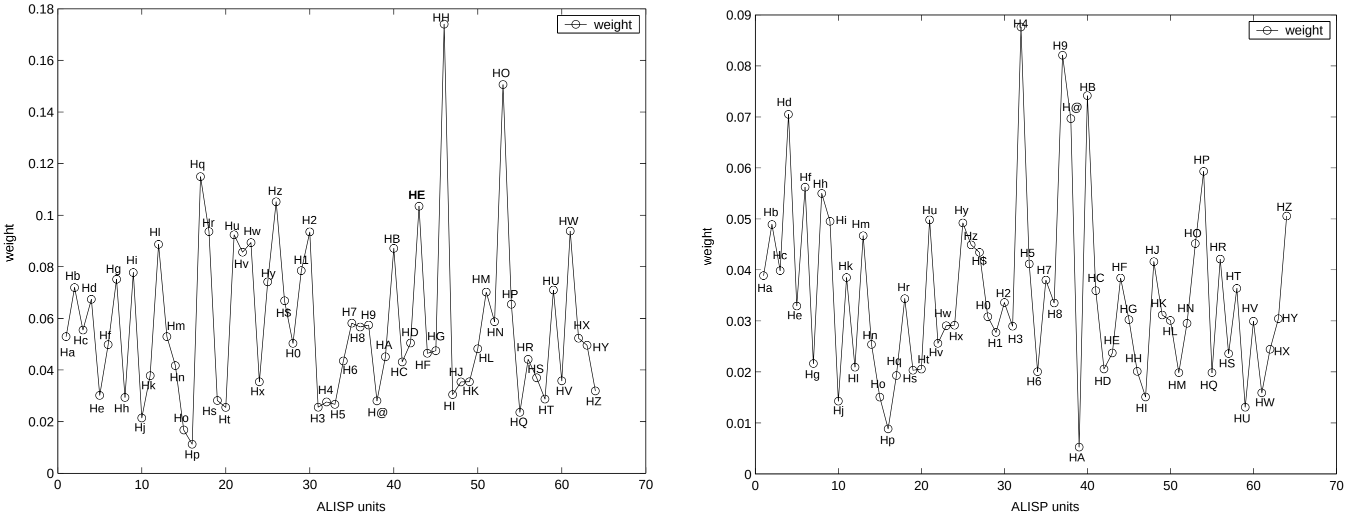
Initially, the proposed system combines the information provided by the DTW distance measure by a simple summation of the segmental score. The final score is obtained by normalizing by the total number of segments in the test file. This gives to each of the segments (belonging to one of the 64 ALIPS speech classes) the same weighting in the overall scoring. This system, whose performance is given in Figure 5 uses all of the segments present to produce the final score. However, as already mentioned in the Introduction, some of the segments provide little or no discrimination between speakers and the inclusion of these may lead to a degradation in performance. Therefore it should be possible to improve the system performance by omitting or reducing the contribution of these segments. An example is the usual removal of frames that are supposed to represent the "silences", and that are usually removed, and not used for the speaker verification procedure.

If we use the Logistic Regression, as explained in Section II to determine the optimal weights for the merging(fusion) of the segmental scores, we achieve a better performance, than the linear summation of the scores. The class and gender dependent weights are estimated form the Development Set. Figure 4 shows the weights, found by the Logistic Regression, for each segment. Because the speech segmentation is gender dependent, the weights are also gender dependent. The distribution of weights confirms that certain segments perform significantly better than others. The results, when we use these wights on the evaluation set are shown in Figure 5. These results show that a significant improvement in performance has been made by weighting the segment classes.

In Figure 5, all the segments are used. In order to have a better representation of which are the worst and best performing classes, we listened to some files. We found that the segments with low weight generally correspond to "silences" and those with high weights (HH for male and H4 for female) correspond to vowels.

In order to reduce the calculation duration, we have applied a threshold of $U = 0.025$ to the weights. The results of the speaker verification performance, when not considering the speech segments that belong to classes with a weight below the threshold U, and using the other weights, are shown on Figure 6, with the solid line. For comparison, the results (dotted lines) including all the segments, with the weights determined by the Logistic Regression are repeated once more.

Using a system which excludes the silence automatically provides better results than the reference system using all of the segments. It should be noted also that these results are



(a) male

(b) female

Fig. 4. Weight for each ALISP unit

obtained without the usual normalization techniques (like Z or T Normalization), commonly used for speaker verification experiments. Further improvement of our system could be foreseen if we use them with the proposed system. For comparison, we can indicate that the equal-error rate for a standard GMM system is about 8% (best single GMM results from 2003 NIST Speaker Recognition Evaluation).

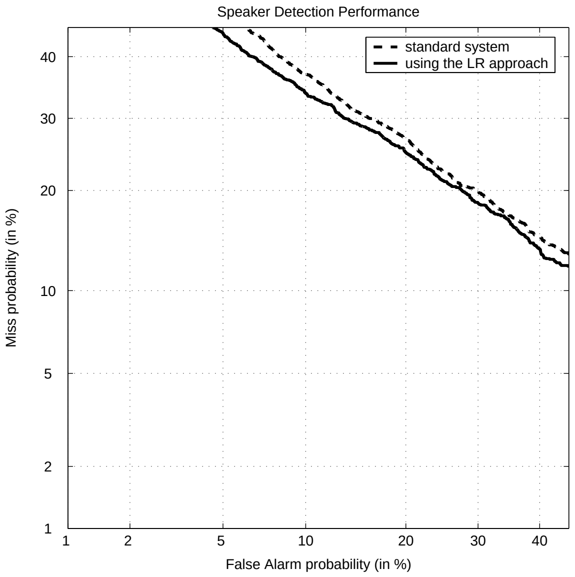


Fig. 5. Speaker verification results for the linear fusion and the fusion using the Logistic Regression approach.

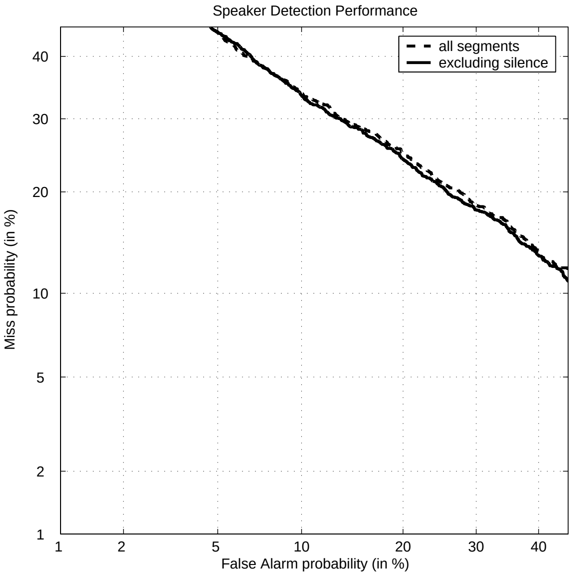


Fig. 6. Performance of the system using all segments and using only the most discriminative segments.

V. Conclusion

In this paper a comparison between two segmental speaker verification systems is presented. For the first one, the scores are calculated for each ALISP segment, and an equal weight is given to each classes. For the second system, we use the Logistic Regression, to determine class specific weights for each ALISP class. The results show that with the Logistic Regression, we can determine the segments that are more discriminant for the speaker verification task, and we can also detect automatically the majority of the segments corresponding to “silences”. Further improvements of our system could be foreseen applying the commonly used speaker verification normalization techniques. Future work will concentrate on traying to fuse our system with a GMM based system.

ACKNOWLEDGMENT

We wish to thank Sophie-Charlotte Barrière and Philippe Froidevaux, ENST and DIVA system managers, for their help. Our thanks go also to F. Bimbot for his tsd95 package, to J. Černocký for the ALISP tools, and to the WaveSurfer and HTK teams.

REFERENCES

[1] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: *Speaker Verification Using Adapted Gaussian Mixture Models*. Digital Signal Processing, Special Issue on the NIST’99 evaluations, Vol. 10(1-3), 19–41, January/April/July 2000

[2] Eatock, J.P., Mason, J.S.: *A Quantitative Assessment of the Relative Speaker Discriminant Properties of Phonemes*. Proc. ICASSP, volume 1, 133–136 (1994)

[3] Olsen, J.: *A Two-stage Procedure for Phone Based Speaker Verification*. In G. Borgefors J. Bigün, G. Chollet, editor, First International Conference on Audio and Video Based Biometric Person Authentication (AVBPA), Springer Verlag: Lecture Notes in computer Science 1206. 199–226 (1997)

[4] Petrovska-Delacrétaz, D., Černocký, J., Hennebert, J., Chollet, G.: *Text-independent Speaker Verification Using Automatically Labeled Acoustic Segments*. In International Conference on Spoken Language Processing (ICLSP), Sydney, Australia (1998)

[5] Petrovska-Delacrétaz, D., Černocký, J., Chollet, G.: *Segmental Approaches for Automatic Speaker Verification*. Digital Signal Processing, Special Issue on the NIST’99 evaluations, Vol. 10(1-3), 198–212, January/April/July 2000

[6] Chollet, G., Černocký, J., Constantinescu, A., Deligne, S., Bimbot, F.: *Towards ALISP: a proposal for Automatic Language Independent Speech Processing*. In Keith Ponting, editor, NATO ASI: Computational models of speech pattern processing Springer Verlag (1999)

[7] Rabiner, L., Schafer, R.W.: *Digital Processing of Speech Signals*. Prentice Hall, Engewood Cliffs, NJ (1978)

[8] Atal, B.: *Efficient coding of LPC Parameters by Temporal Decomposition*. Proc. IEEE ICASSP 83, 81–84 (1983)

[9] Reynolds, D.A.: *Comparison of Background Normalization Methods for Text-independent Speaker Verification*. Proc. Eurospeech,

[10] Reynolds, D.A.: Comparison of Background Normalization Methods for Text-independent Speaker Verification. Proc. Eurospeech, Rhodes, 963–966 (1997)

[11] Rosenberg, A. E.: Automatic Speaker Verification: A Review. Proc. IEEE, Vol. 64, No. 4, pp. 475-487, April 1976.

[12] Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 29, No. 2, pages 254-272, 1981.

[13] Pandit, M., Kittler J.: Feature Selection for a DTW-Based Speaker Verification System. Proc. ICASSP, Seattle, Vol. 2 (1998) 769-772

[14] S. Pigeon, P. Druyts, and P. Verlinde, *Applying Logistic Regression To The Fusion Of The Nist’99 1-Speaker Submissions*. Digital Signal Processing, 10(1):237-248, January/April/July 2000.

[15] Hosner, D.W. and Lemeshow, S., *Applied Logistic Regression*. Wiley, New York, 1989.

[16] Verlinde, P. and Chollet, G., *Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multimodal identity verification application*. In Proceedings of the Second International Conference on Audio- and Video-Based Biometric Person Authentication, Washington, DC, March 1999, pp. 188-193.

[17] D. Petrovska-Delacretaz, Asmaa El Hannani, G. Chollet.: *Searching through a speech memory for text-independent speaker verification*. In proc. of Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA). Guildford (UK), June 2003

[18] A. Martin and M. Przybocki, *The NIST Year 2002 Speaker Recognition Evaluation Plan*, `http://www.nist.gov/speech/tests/spk/2002/doc/2002-spkrec-evalplan-v60.pdf`

An Audio-Visual Person Identification and Verification System Using FAPs as Visual Features

Petar S. Aleksic and Aggelos K. Katsaggelos

Department of Electrical and Computer Engineering
Northwestern University
2145 N. Sheridan Road, Evanston, IL 60208
Email: {apetar, aggk} @ece.northwestern.edu

Abstract

The performance and robustness of a person recognition system using acoustic information can be improved with the use of visual information. In this paper we present an audio-visual (AV) speaker identification and verification system and analyze its performance. The proposed system utilizes a Hidden Markov Model (HMM) approach and Facial Animation Parameters (FAPs), supported by the MPEG-4 standard for representing visual features. A number of experiments have been performed under both clean, and noisy (utilizing additive Gaussian noise) audio conditions at different signal-to-noise ratios (SNRs) ranging from 0 dB to 20 dB. The proposed system improves the performance of the audio-only system at all SNRs tested and under clean audio conditions.

1. Introduction

Person recognition can be classified into two problems, person identification and person verification [1]. Person identification is the problem of determining the identity of a person (who the person is) from a closed set of candidates, based on the best match of the person's biometric signals to those in a database. Person verification is the problem of determining whether a person is whom s/he claims to be, also utilizing the person's biometric signals. A verification system should be able to reject claims from impostors, persons not registered with the system, and accept claims from the clients, persons registered with the system. There is an increasing need of reliable person verification systems to improve the security of systems or services used by only selected group of people. There are many different biometric signals, such as faces, voices, fingerprints, iris scans, and passwords, that can be used in a person recognition system [2, 3]. Each modality has its own advantages and disadvantages. Although single modality biometric systems can achieve high performance in some cases, they are usually not robust to noise and do not meet the needs of many potential person verification (or recognition) applications. It has been shown that using multiple biometric modalities instead of a singe modality improves the performance of a system [4, 5]. Different modalities are combined in order to eliminate problems characteristic for single modalities. For example, a person's voice and face, as biometric signals, are easily collected and natural to the user. Person verification systems that rely only on audio data are sensitive to acoustic noise and therefore not acceptable for many high security applications. On the other hand, systems that rely only on visual data can also be very sensitive to visual noise (lightning changes, poor video quality, occlusion, etc.). Audio and visual data have been used in automatic speech recognition (ASR) applications in order to improve ASR performance [6, 7]. The fact that the information present in the visual signal can be used not only to improve speech recognition performance but also to characterize a persons' identity justifies the use of audio-visual biometric systems for person recognition applications.

An important factor in designing an audio-visual recognition system is the selection of the audio and visual features. While the selection of audio features is a well-studied and agreed upon issue, various visual features have been utilized. MPEG-4 is an audiovisual object-based video representation standard supporting facial animation. MPEG-4's facial animation is controlled by the Facial Definition Parameters (FDPs) and Facial Animation Parameters (FAPs), which describe the face shape, and movement, respectively [8]. The MPEG-4 standard defines 68 FAPs, divided into 10 groups (group 8 FAPS pertaining to the mouth area –an important visual speech articulator- are shown in Figure 1). Transmission of all FAPs at 30 frames per second requires only around 20 kbps (or just a few kbps, if MPEG-4 FAP interpolation is efficiently used [9]), which is much lower than standard video transmission rates. FAPs represent an important descriptor of visual articulatory information whish is, clearly standard-compliant, and also portable, i.e., different 3D facial models [10] can be animated successfully by the same stream of FAPs. They are therefore utilized in the proposed system.

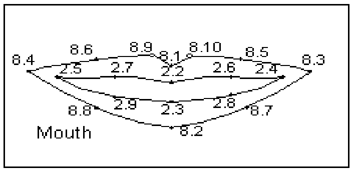
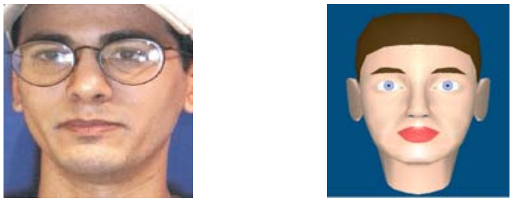


Figure 1. Facial animation parameters (FAPs)



a) b)

Figure 2. a) Original video frame; b) MPEG-4 model [10]

In this paper we describe an audio-visual person recognition system that uses Hidden Markov Models (HMMs) to model the temporal behavior of audio and visual data. Word-level continuous HMMs are used to model the temporal statistics of audio and visual data. Each person in the database is modeled using a separate HMM. During the training procedure, the world model is first trained on the training data of all speakers. The world model is then used as the initial model for each speaker HMM, which is retrained using only the training part of the database corresponding to the particular speaker.

To the best of our knowledge no results have been previously reported in the literature on AV identification and verification when FAPs are used as visual features. It is therefore the main objective of this paper to report on such results.

In this paper, we first describe in Section 2 the visual features extraction and in Section 3 the audio-visual integration approach used. Next the person verification and identification experiments and training procedures are described in Section 4. Finally, conclusions are drawn and future work is proposed in Section 5.

2. Visual feature extraction

This work utilizes the CMU (Carnegie Melon University) audio-visual database [11]. The database contains ten subjects; seven of whom are male and three female. The vocabulary includes 78 words commonly used for scheduling applications. Each subject repeated each of the words ten times. For each of the word set repetitions, the database contains a speech waveform and a word-level transcription. The video (a sample frame is shown in Figure 2a) was sampled at a rate of 30 frames per second while audio was acquired at a rate of 16 kHz. All FAPs are expressed in terms of Facial Animation Parameter Units (FAPU), as shown in Figure 3. These units are normalized by important facial feature distances in order to give an accurate and consistent representation. Ten FAPs from group 8 which describe the outer lip contours were used in our work, and they are represented with the use of two FAPUs, mouth-width separation (MW) and mouth-nose separation (MNS). Each of these two distances is normalized to 1024.

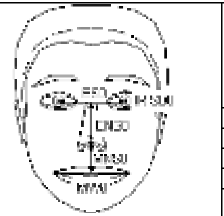
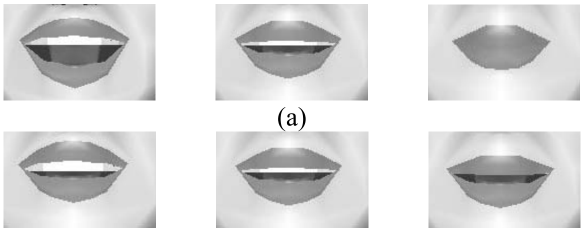
Description		FAPU value
	IRIS Diameter (by definition it is equal to the distance between upper and lower eyelid) in neutral face	IRISD = IRISD0 / 1024
	Eye Separation	ES = ES0 / 1024
	Eye - Nose Separation	ENS = ENS0 / 1024
	Mouth - Nose Separation	MNS = MNS0 / 1024
	Mouth - Width Separation	MW = MW0 / 1024
	Angular Unit	AU = 10^{-5} rad

Figure 3. Facial Animation Parameter Units (FAPU)



(a)

(b)

Figure 4. The mean lip shape (middle column), and the lip shapes obtained by the variation of the first(a) and second (b) eigenvector weights by +2 standard deviations (left column) and –2 standard deviations (right column)

The FAP sequences were extracted from the available visual data for all ten subjects and for all 100 word sequences. Through visual evaluation of the FAP extraction results we observed that the extracted parameters produced a natural movement of the MPEG-4 decoder (Figure 2b) that synchronized well with the audio.

In order to decrease the dimensionality of the visual feature vector, Principal Component Analysis (PCA) was performed on the 10-dimensional FAP vectors ($\boldsymbol{f}_n$). The PCA training set consists of *N* FAP vectors, which are obtained from the training part of the visual data. The *10x10* covariance matrix $\boldsymbol{C}$ can be computed as

$$\boldsymbol{C} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{f}_n - \bar{\boldsymbol{f}}_n)(\boldsymbol{f}_n - \bar{\boldsymbol{f}}_n)^T , \qquad (1)$$

where $\bar{\boldsymbol{f}}_n$ denotes the mean FAP vector.

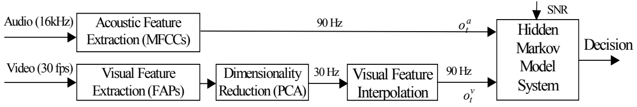

Audio (16kHz)
Acoustic Feature Extraction (MFCCs)
90 Hz
o_t^a
SNR
Hidden Markov Model System
Decision
Video (30 fps)
Visual Feature Extraction (FAPs)
Dimensionality Reduction (PCA)
30 Hz
Visual Feature Interpolation
90 Hz
o_t^v

Figure 5. Audio-visual system for ASR

After the covariance matrix was obtained and its eigenvalues determined, the FAPs, $\boldsymbol{f}_n$, were projected onto the eigenspace defined by the first K eigenvectors,

$$\boldsymbol{f}_n = \bar{\boldsymbol{f}}_n + \boldsymbol{E} \cdot \boldsymbol{o}_n^f , \qquad (2)$$

where $\boldsymbol{E}$=[$\boldsymbol{e}_1$ $\boldsymbol{e}_2$...$\boldsymbol{e}_K$] is the matrix of K eigenvectors, which correspond to the K largest eigenvalues, and $\boldsymbol{o}_n^f$ the $Kx1$ vector of corresponding projection weights. The first six, three, and one eigenvectors represent respectively 99%, 95%, and 82% of the total statistical variance. By varying the projection weights by ±2 standard deviations, we concluded that the first and second eigenvectors mostly describe the movement of the lower and upper lip, respectively (Figure 4), while the third eigenvector mostly describes mouth shape asymmetries. When choosing the dimensionality of the visual feature vector it should be kept in mind the trade-off between the number of HMM parameters that have to be estimated and the amount of the person recognition information contained in the visual features. Based on the statistical variance distribution and the above-mentioned trade-off we decided to use three-dimensional ($K=3$) projection weights as visual features. These features were used in all audio-visual person verification and identification experiments we conducted.

3. Audio-visual integration (feature fusion)

In the system we developed the audio and visual streams are combined as shown in Figure 5. The Mel-Frequency Cepstral Coefficients (MFCC), signal energy and first and second derivatives, widely used in speech processing, were used as audio features. The three-dimensional projections weights (Eq. 3) and their first and second derivatives were used as visual features. The size of the audio features was 39. The size of the visual features was 9. Since audio features (MFCCs) were extracted at a rate of 90Hz, while visual features (FAPs) at a rate of 30Hz, the visual features were interpolated in order to obtain synchronized data.

In this approach the audio-visual feature observation vector ($\boldsymbol{o}_t$) is formed by appending the visual observations vector ($\boldsymbol{o}_t^v$) to the audio observations vector ($\boldsymbol{o}_t^a$), that is

$$\boldsymbol{o}_t = [\boldsymbol{o}_t^{a\,T} \quad \boldsymbol{o}_t^{v\,T}]^T . \qquad (3)$$

The newly obtained joint features vectors were used to train an HMM model, with continuous state emission probabilities [12] given by

$$\boldsymbol{b}_j(\boldsymbol{o}_t) = \sum_{m=1}^{M} c_{jm} N(\boldsymbol{o}_t ; \boldsymbol{\mu}_{jm} , \boldsymbol{\Sigma}_{jm}) \qquad (4)$$

In Eq. 4 subscript j denotes a state of a word model, M denotes the number of mixtures, c_{jm} denotes the weight of the m'th mixture component, and N is a multivariate Gaussian with mean vector $\boldsymbol{\mu}_{jm}$ and diagonal covariance matrix $\boldsymbol{\Sigma}_{jm}$. The sum of mixture weights c_{jm} is equal to 1.

4. Training and experimental results

The baseline HMM system was developed using the HTK toolkit version 3.1 [13]. In this work a text-dependent audio-visual person verification and identification system is considered. The experiments used the portion of the CMU database in which speakers utter the digit sequence "234567". The HMMs used to model each speaker had left-right topology, with 10 states. The part of the CMU audio-visual database used in the experiments consists of each speaker uttering the digit sequence 10 times. The data is divided into training, evaluation, and testing parts. The first six utterances of each speaker were used for training, one utterance was used for evaluation, and the remaining three for testing. The same training and testing procedures were used for both audio-only and audio-visual experiments. To test the algorithm over a wide range of SNRs (0, 10, 20 dB), white Gaussian noise was added to the audio signals. All results were obtained using HMMs trained in matched conditions, by corrupting the training data with the same level of noise, as used for corrupting the testing data. This approach was used in order to accurately measure the influence of the visual data on system performance.

Word-level continuous HMMs were trained for each speaker in the database. During the training procedure, the world model (M_W) is first trained on the training data of all speakers. The world model is used as the initial model for each speaker HMM (M_S). Each speaker HMM is then retrained using only the training part of the database corresponding to the particular speaker.

In the person identification experiments the objective was to determine the speaker ($\hat{s}$) who's HMM matches the best the unknown person's data ($\boldsymbol{o}_t$), that is

$$\hat{s} = \arg\max_{s \in S} \Pr(M_s \mid \mathbf{o_t}) , \qquad (5)$$

where S denotes set of all speakers in the database, and M_s an HMM for speaker s.

In the person verification experiments the objective was to accept the client claims and reject impostor claims. The similarity measure (D_{HMM}) is defined as the likelihood ratio between the speaker set and the world set, that is

$$D_{HMM} = \log \Pr(M_s \mid \mathbf{o_t}) - \log \Pr(M_W \mid \mathbf{o_t}) . \qquad (6)$$

If the similarity measure is larger than the *a priori* defined verification threshold the claim is accepted, and otherwise it is rejected. The evaluation part of the data was used to calculate the verification thresholds to be used in determining whether a person is accepted or rejected. The thresholds determined from the evaluation set were used for testing.

Table 1. Person identification results

Person Identification Error [%]		
SNR [dB]	**Audio only**	**Audio-visual**
clean	5.13	5.13
20	19.51	7.69
10	38.03	10.26
0	53.10	12.82

Table 2. Person verification results

SNR [dB]	**Audio only [%]**			**Audio-visual [%]**		
	FA	**FR**	**EER**	**FA**	**FR**	**EER**
clean	2.85	25.64	2.56	0	12.82	1.71
20	2.85	41.03	3.99	2.85	20.51	2.28
10	0	53.85	4.99	0	23.08	2.71
0	5.7	61.54	8.26	2.85	28.21	3.13

Two commonly used error measures for a verification system are False Acceptance (FA) –an impostor is accepted - and False Rejection (FR) –a client is rejected. They are defined by

$$FA=\frac{I_A}{I}\times 100\% \quad FR=\frac{C_R}{C}\times 100\%\,, \qquad (7)$$

where I_A denotes the number of accepted impostors, I the number of impostor claims, C_R the number of rejected clients, and C the number of client claims. There is a trade-off between FA and FR, which is controlled by the *a priori* chosen verification threshold. Verification system performance can also be measured using Equal Error Rate (EER). It is determined after the verification experiments are performed, by choosing the threshold for which FA and FR are equal.

The verification threshold is chosen on the evaluation set to meet certain FA and FR requirements. In our experiments we set the threshold to obtain the minimum FA rate. FA and FR rates for that threshold are shown in Table 2. The threshold for which FA and FR were equal is also calculated after all the verification experiments were performed in order to determine EER. The EER results are also shown in the Table2.

The results of the person identification experiments obtained for different levels of acoustic noise for both audio-only and audio-visual approaches are shown in Table 1. The audio-visual identification system outperforms the audio-only system for all SNRs tested and achieves the same performance as the audio-only system under clean audio conditions.

The results of the person verification experiments are shown in Table 2. As can be clearly seen, the performance of the audio-only system degrades significantly in the presence of noise. The proposed audio-visual system performs considerably better than the audio-only system for all SNRs and for the clean speech. It is important to point out that the considerable performance improvement was achieved, although only nine-dimensional visual features were used.

5. Conclusions

We have described an audio-visual person verification and identification system that significantly improves performance over an audio-only system. Our system uses FAPs, supported by the MPEG-4 standard for the visual representation as visual features. We plan to extract additional FAPs and determine how much information useful for person recognition is contained in them. We also plan to perform text-independent experiments and experiments on larger AV database.

6. References

[1] J. Luettin, "Speaker Verification Experiments on the M2VTS Database," IDIAP-RR 99-02, 1999.

[2] *J. Luettin, N. A. Thacker, and S. W. Beet*, "Speaker identification by lipreading," in *Proceedings of the 4th Int. Conference on Spoken Language Processing (ICSLP'96)*, 1996.

[3] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," to appear in *IEEE Transactions on circuits and Systems for Video Technology, Special Issue on Image- and Video-Based Biometrics*, August 2003.

[4] S. Ben-Yacoub, Y. Abdeljaoued and E. Mayoraz, " Fusion of face and speech data for person identity verification," *IEEE Trans. On Neural Networks*, Vol. 10, No. 5, 1999, pp. 1065-1074.

[5] C. Sanderson and K. K. Paliwal. Information Fusion and Person Verification Using Speech & Face Information. IDIAP Research Report 02-33, Martigny, Switzerland, 2002.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audio-visual speech," to appear: *Proc. IEEE*, 2003.

[7] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features", *EURASIP Journal on Applied Signal Processing,* pp.1213-1227, 2002.

[8] Text for ISO/IEC FDIS 14496-2 Visual, ISO/IEC JTC1/SC29/WG11 N2502, Nov. 1998.

[9] F. Lavagetto and R. Pockaj, "An efficient use of MPEG-4 FAP interpolation for facial animation at 70 bits/frame," *IEEE Trans. on Cir. and Sys. for Video Tech.*, Vol. 11(10), pp.1085-1097, October 2001.

[10] G. A. Abrantes, FACE-Facial Animation System, version 3.3.1, Instituto Superior Tecnico, (c) 1997-98.

[11] http://amp.ece.cmu.edu/

[12] L. Rabiner, B.-H. Juang, “Fundamentals of speech Recognition,” Prentice Hall, Englewood Cliffs, 1993.

[13] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “The HTK Book,” Entropic Ltd., Cambridge, 1999.

Audio-Visual Person Authentication Using Speech and Ear Images

Koji Iwano, Tomoharu Hirose, Eigo Kamibayashi, and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{iwano, hirose, kamibaya, furui}@furui.cs.titech.ac.jp

Abstract

This paper proposes a multimodal, biometric person authentication method using speech and ear images to attempt to improve the performance in mobile environments. It is well known that the performance of person authentication using only speech is deteriorated by acoustic noises and feature changes with time. Since the ear shape of each person does not change over time, integrating its image with speech information increases robustness of person authentication. Experiments are conducted using audio-visual database collected from 38 male speakers at five sessions over a half year period. Speech data are contaminated with white noise at various SNR conditions. Experimental results show that the authentication performance is improved by combining the ear image with speech in every SNR condition.

1. Introduction

The necessity of person authentication is spreading in the recent network society. Biometric authentication, which identifies an individual person using physiological and/or behavioral characteristics, such as face, fingerprints, hand geometry, handwriting, iris, retinal, vein, and speech, is one of the most attractive and effective methods. These methods are more reliable and capable than knowledge-based (e.g., password) or token-based (e.g., a key) techniques, since biometric features are hardly stolen or forgotten.

Although “speech” is one of the most useful and effective features for person authentication in mobile environments, its performance deteriorates due to additive noise and session-to-session variability of voice quality. Therefore, the combination with other biometric features to improve the performance has attracted a great deal of attention. Along this line, various audio-visual biometric authentication methods have been proposed[1, 2, 3, 4, 5]. Although most of them use “face” information in combination with speech, the face features also change due to make-up, mustache, beard, hair styles and so on, and derives degradation of the performance. Therefore, it is worth investigating other biometric features with high permanence.

From this point of view, this paper proposes an authentication method using “ear” shape information in combination with speech. It is well known that the ear shape hardly changes over time[6, 7]. Although several authentication methods using ear images have already been proposed[7, 8, 9], there is no research on multimodal authentication using both speech and ear images. Since ear images could be captured using a small camera installed in a mobile phone, ear information is expected to be easily used in mobile environments than other biometrics, such as fingerprint, iris, and retinal, that need special equipment.

Our authentication method and audio-visual database are described in Section 2. Section 3 reports experimental results and Section 4 concludes this paper.

2. System structure and experiments

Figure 1 shows our multimodal person authentication system using speech and ear images. Audio and visual data are respectively converted into feature vectors. Each set of features is matched with both a claimed person model and a speaker independent (SI) model. Then, audio and visual scores are integrated with appropriate weighting and a decision is made whether he/she is a true speaker or an impostor. If the score is larger than a threshold value, the speaker is accepted as a claimed speaker.

2.1. Integrated score

A posterior probability is used as the authentication score. The posterior probability of being a claimed speaker S^c after observing a biometric feature set x, is denoted by $p(S^c|x)$. Since x is composed of speech (audio) features x_s and ear (visual) features x_e, $p(S^c|x)$ can be transformed as follows:

$$p(S^c|x) = p(S^c_s|x_s) \cdot p(S^c_e|x_e) \qquad (1)$$

where S^c_s and S^c_e represent the claimed speaker’s speech and ear models, respectively. Bayes’ Rule derives the following

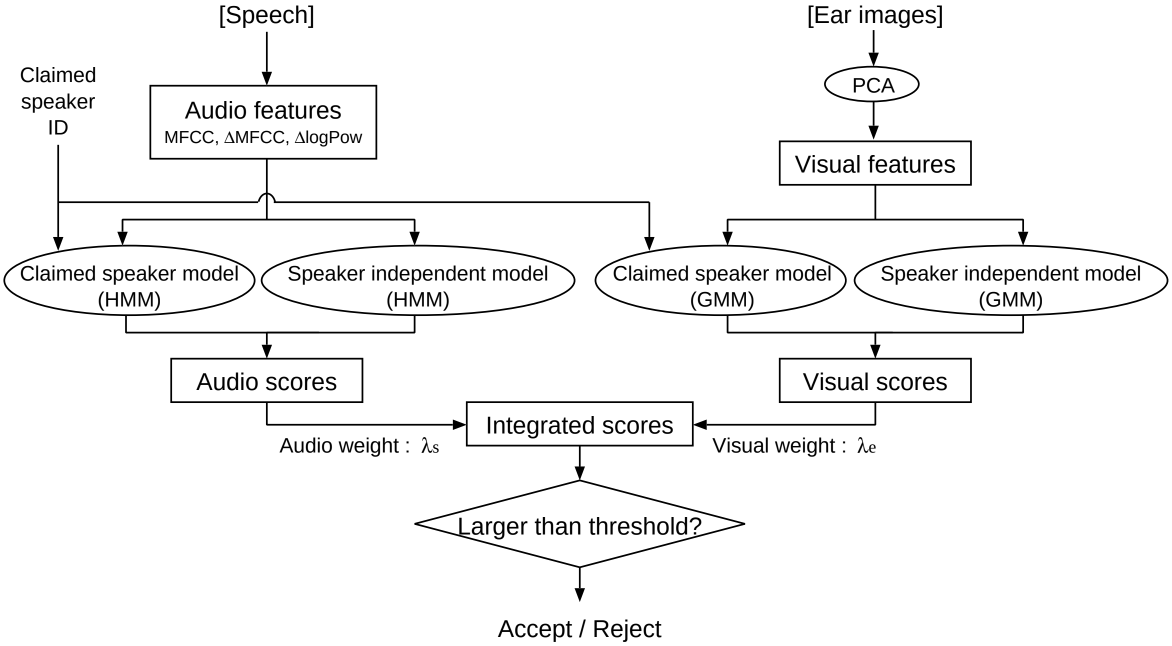


Fig. 1. Multimodal person authentication system using speech and ear images.

equation:

$$p(S^c|x) = \frac{p(x_s|S_s^c)p(S_s^c)}{p(x_s)} \cdot \frac{p(x_e|S_e^c)p(S_e^c)}{p(x_e)} \qquad (2)$$

where $p(x_s|S_s^c)$ and $p(x_e|S_e^c)$ are likelihood values with claimed speaker's speech and ear models, respectively. The probabilities in the denominator are approximated by using likelihood values with general speaker's speech model $p(x_s|S_s^g)$ and ear model $p(x_e|S_e^g)$:

$$\begin{aligned} p(S^c|x) &\approx \frac{p(x_s|S_s^c)p(S_s^c)}{p(x_s|S_s^g)p(S_s^g)} \cdot \frac{p(x_e|S_e^c)p(S_e^c)}{p(x_e|S_e^g)p(S_e^g)} && (3) \\ &\propto \frac{p(x_s|S_s^c)}{p(x_s|S_s^g)} \cdot \frac{p(x_e|S_e^c)}{p(x_e|S_e^g)} && (4) \end{aligned}$$

Equation (4) means that the posterior probability for the claimed speaker's model is calculated by the product of likelihood values normalized using speaker independent (SI) models. By defining authentication scores for speech (p_s) and ear (p_e) as

$$p_m = \log p(x_m|S_m^c) - \log p(x_m|S_m^g) \;\; (m = s, e) \qquad (5)$$

an integrated score p_{se} which balances the effectiveness of speech and ear features can be modeled by the following equation.

$$p_{se} = \lambda_s p_s + \lambda_e p_e \;\; (\lambda_s + \lambda_e = 1) \qquad (6)$$

where λ_s and λ_e are audio and visual weights, respectively.

2.2. Audio-visual database

2.2.1. Recording conditions

Audio-visual data were recorded at five sessions with intervals of approximately one month. The data were collected from 38 male speakers, and each speaker uttered 50 strings of four connected digits in Japanese at each session. Speech data were sampled at 16kHz with 16bit resolution. One right ear image for each speaker taken by a digital camera with 720×540 pixel resolution was collected at each session. Figure 2 shows the arrangement of a speaker and a camera when recording. An image of the whole ear, with no hair obscuring it, was captured by the camera positioned perpendicular to the ear. The camera was located approximately 20cm away from each speaker's ear. A flash was used to keep constant illumination.

2.2.2. Training and testing data

A set of data recorded at sessions 1∼3 was used for training and that recorded at sessions 4 and 5 was used for testing. The database was separated into two groups in terms of speakers as shown in Figure 3. This figure shows the case that the speaker #01 was used as the claimed speaker. The SI model was trained using the utterances by all the speakers in the speaker group B which did not include the claimed speaker. When one of the speakers in the speaker group B was used as the claimed speaker, the utterances by

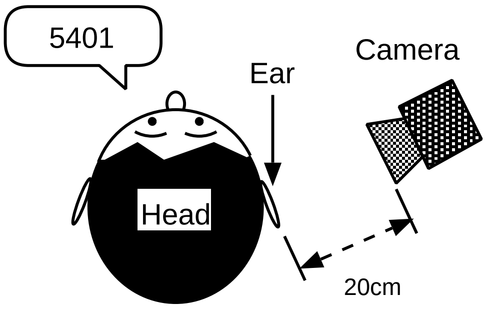


Fig. 2. Location of speaker and camera.

	Trainig data	Testing data	
Speaker ID	Session 1,2,3	Session 4,5	
#01 : : #19	Uesd for speaker model	True speaker Impostors	Group A
#20 : : #38	Used for speaker independent model	Impostors	Group B

Fig. 3. Training and testing data for the authentication experiment when the speaker #01 is the claimed speaker.

the speaker group A were used for the SI model training. In this way, the SI model was always trained using the data of a speaker group not including the claimed speaker. All the speakers in both speaker groups A and B, except for the claimed speaker himself, were used as imposters.

White noise was added to the audio data for training at 30dB SNR level to increase the robustness against noisy speech, and testing data were contaminated with white noise at 5, 10, 15, 20, and 30dB SNR conditions.

As image data, we first extracted gray-scaled ear images with 80×80 pixel resolution. An example of the extracted ear image is shown in Figure 4. The ear location and rotation in the image were manually adjusted. In order to increase robustness of visual models, the following variations were given to training data:

(1) Shifting the ear location in vertical and horizontal directions within ±6 pixels at a 2 pixel interval. Consequently, 49 variations were made for each ear image.

(2) Rotating the ear images within ±30 degrees at one degree interval. Accordingly, 61 variations were made for each ear image.

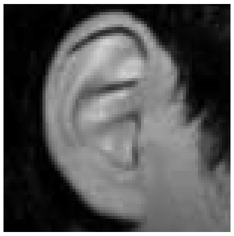


Fig. 4. An example of the extracted ear image.

The both operations made approximately 9,000 (= 3 sessions × 49 × 61) ear images for training each speaker's model. For testing data, we applied only the rotating operation (2).

Both training and testing data were filtered to emphasize the ear feature. The following three conditions were experimentally compared to find the best filtering method:

(a) No filtering (Figure 5(a)).

(b) Laplacian filtering (Figure 5(b)).

(c) Laplacian-Gaussian filtering (Figure 5(c)).

Finally, all ear images were circularly sampled and digitized for reducing hair effects and avoiding the window shape effects caused by rotation of the images.

2.3. Audio and visual features

Audio features were 25-dimensional vectors consisting of 12 MFCCs, 12ΔMFCCs, and Δ log energy. The frame shift was 10ms and the analysis window length was 25ms. For ear images, "eigen-ear" space was built by using Principal Components Analysis (PCA) in the same way as the eigenface approach used in face recognition[10]. The PCA was applied to the ear images recorded at the first session using 19 speakers in one of the two speaker groups that did not include the claimed speaker. The original ear images with no shifting or rotating were used for the analysis. Figure 6 shows examples of the first eight eigen-ear images obtained by the PCA using the Laplacian-Gaussian filtered images. All the ear images were converted into 18-dimensional visual feature vectors using the first 18 eigen-ears.

2.4. Speech and ear models

The audio features were modeled by digit-unit HMMs. Each digit HMM has a standard left-to-right topology with $n \times 3$ states, where n is the number of phonemes in the digit. The

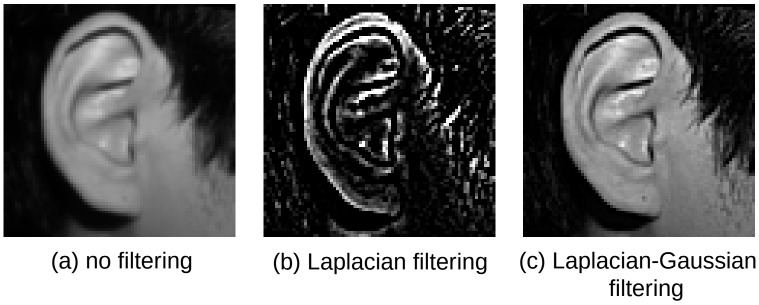


Fig. 5. Examples of the filtered ear images.

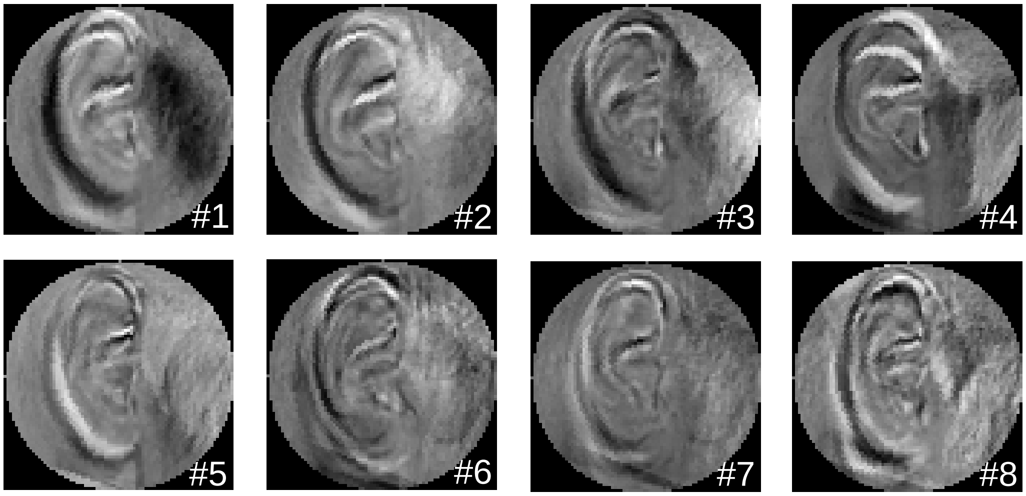


Fig. 6. Examples of the first 8 eigen-ear images.

authentication score for the speech features represented in Equation (4) is calculated as follows:

$$\begin{aligned}\frac{p(x_s|S_s^c)}{p(x_s|S_s^g)} &= \frac{\Sigma_w p(x_s|S_s^c, w)p(w)}{\Sigma_w p(x_s|S_s^g, w)p(w)} \\ &\approx \frac{\max_w p(x_s|S_s^c, w)}{\max_w p(x_s|S_s^g, w)} \qquad (7)\end{aligned}$$

where w is a string of four connected digits.

The visual features were modeled using GMMs. In each testing experiment, 61-feature vectors converted from the rotated images were input to the GMMs. Log likelihood values calculated for the claimed speaker and the SI models were used to obtain the authentication score for each ear image according to the Equation (5).

3. Experimental results

3.1. Results of the authentication using ears

An experiment using only ear images was first conducted for investigating the effects of shifting and filtering the ear images. Table 1 shows equal error rates (EER) for the person authentication at various conditions of filtering and image processing applied to the training data. In the experiment, optimum numbers of mixtures: eight mixtures for speaker GMMs and one mixture for SI GMM, were experimentally chosen.

The results show that both filtering methods are effective for improving the authentication performance. The Laplacian-Gaussian filtering yields better results than the Laplacian filtering. The shifting operation for training data also improves the performance irrespective of filtering methods. This probably means that there are some mismatches of ear location between training and testing data due to the manual image extraction process.

Table 1. Equal error rate (%) in person authentication using ear images with various kinds of filtering and processing in the training stage.

	only rotating	shifting & rotating
no filtering	14.5	14.0
Laplacian filtering	13.6	13.3
Laplacian-Gaussian filtering	13.2	**11.9**

The best result, 11.9% EER, is observed at the condition using the Laplacian-Gaussian filtering and shifting as well as rotating operations. This condition is used in the following visual authentication experiments.

3.2. Results of the multimodal authentication

Multimodal authentication results in various SNR conditions obtained by using optimum audio weights (λ_s) are shown in Figure 7. The optimum weights (λ_s) were determined experimentally to minimize the error rate at each condition. The optimum values are also shown in the figure. Results using only speech ($\lambda_s = 1.0$) and only ear ($\lambda_s = 0.0$) are also shown for the purpose of comparison. The number of mixtures in audio HMMs was optimized based on the experimental results at the 30dB SNR condition; the number of mixtures was set to four for both speaker and SI HMMs.

Although the authentication performance using only speech is highly degraded by the noise effect, it is clearly shown that multimodal authentication is robust. The proposed method is most effective when the SNR is 15dB; the error rate is reduced by 53.0% from the audio only method and 43.9% from the visual only method. The best performance of 0.3% EER is observed at the 30dB SNR condition.

Figure 8 shows EER as a function of the audio weight (λ_s). Improvement using the ear images is observed over a wide range of λ_s. It is also shown that the optimum λ_s values exist in the range of $0.6 \sim 0.8$ at all the noise conditions with the exception of the 5dB SNR condition. This means that the proposed multimodal method is not sensitive to the change of weights and the weight can be easily optimized.

3.3. Comparing ears with faces as biometrics

We previously conducted person authentication experiments using speech and face features[5] in the similar way as that described in this paper. Although the speech and face database has the same number of speakers and recording sessions as the speech and ear database, 38 male speakers and 5 sessions, actual speakers are different between the two databases.

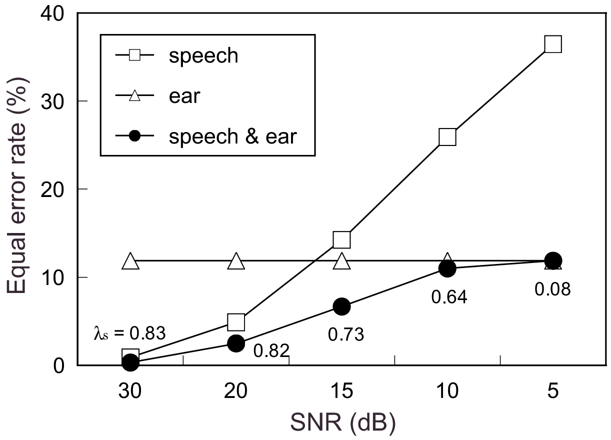


Fig. 7. Person authentication results in various SNR conditions.

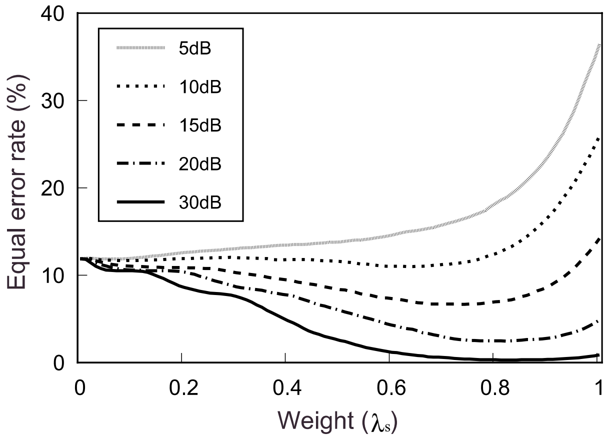


Fig. 8. Equal error rate as a function of the audio weight λ_s.

The previous work showed that the EER using only the face information was 7.0%, which was better than the EER using the ear information, 11.9%.

One of the reasons is that ear images are more changeable than face images by a tilt of the camera, since the ear surface is more irregular than the face surface. However, since the ear itself is not as changeable as the face, the authentication using ear biometrics has a possibility to become a practical method, if the above observation problem can be solved.

4. Conclusions

This paper has proposed a multimodal authentication method using the combination of speech and ear images with the aim of increasing noise robustness in mobile environments. The proposed method has been confirmed to be more robust than the speech only method in various SNR conditions.

Future works include 1) improving the authentication performance using the ear information by increasing the robustness against ear image variation caused by a tilt of a camera, 2) reducing the effects of hair and sideburns, 3) developing an automatic method for ear area detection, and 4) investigating the robustness of ear features against their changes over time.

5. Acknowledgements

This research has been conducted in cooperation with NTT DoCoMo. The authors wish to express thanks for their support.

6. References

[1] R. Brunelli and D. Falavigna, "Personal identification using multiple cues," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.17, no.10, pp.955–966, Oct. 1995.

[2] B. Duc, E.S. Bigun, J. Bigun, G. Maitre, and S. Fischer, "Fusion of audio and video information for multi modal person authentication," *Pattern Recognition Letters*, vol.18, no.9, pp.835–843, Sept. 1997.

[3] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol.18, no.9, pp.853–858, Sept. 1997.

[4] N. Poh, and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Audio- and Video-Based Biometric Person Authentication, Third International Conference, AVBPA 2001*, J. Bigun and F. Smeraldi, Eds., pp.348–353, Springer, 2001.

[5] T. Hirose, K. Iwano, and S. Furui, "Multi-modal speaker verification using speech and face images," *Proc. ASJ Spring Meeting 2003*, vol.1, pp.107-108, March 2003. (In Japanese)

[6] A. Iannarelli, *Ear Identification*. Forensic Identification series. Paramont Publishing Company, Fremont, California, 1989.

[7] M. Burge and W. Burger, "Ear biometrics," in *Biometrics: Personal Identification in Networked Society*, A. Jain, R. Bolle, and S. Pankanti, Eds., pp.273–285, Kluwer Academic, Boston, MA, 1999.

[8] N. Tashiro, K. Shinohara, M. Abe, and T. Okamura, "Individual identification by outline of components on pinna and superposition," *ITE Tech. Rep.*, MIP2001-54, vol.25, no.22, pp.7–13, March 2001. (in Japanese)

[9] Y. Wang, K. Takeda, K. Sato, and S. Nakayama, "Study on human recognition by ear image with eigenear," *Tech. Rep. of IEICE*, IE2002-95, vol.26, no.76, pp.37–42, Nov. 2002. (in Japanese)

[10] M. Turk and A.P. Pentland, "Face recognition using eigenface," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp.586–591, 1991.

Human Ear Recognition in 3D

Bir Bhanu and Hui Chen
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{bhanu,hchen}@vislab.ucr.edu

Abstract

The scant research with ear biometric has used intensity images and therefore the performance of the systems is greatly affected by imaging problems such as lighting and shadows. Range sensors which are insensitive to above imaging problems can directly provide us 3D geometric information. In this paper, we present a new local surface descriptor for surface representation for recognizing human ears in 3D. A local surface descriptor is defined by a centroid, its surface type and 2D histogram. The 2D histogram consists of shape indexes, calculated from principal curvatures, and angles between the normal of reference point and that of its neighbors. By comparing local surface patches between a test and a model image, we find the potential corresponding local surface patches. Geometric constraints are used to filter the corresponding pairs. Verification is performed by estimating transformation and aligning models with the test image. Experiment results with real ear range image are presented to demonstrate the effectiveness of our approach.

1. Introduction

Faces and fingerprints are popular biometrics for personal identification. However, they have some drawbacks. For instance, it's a very challenging problem to design face recognition techniques which can deal with the effects of aging, facial expressions and problems such as changing 3D pose, lighting and shadow. Fingerprints also require the cooperation of subjects. The ear, which is viable as a biometric [13], has certain advantages over other biometrics. For example, ear is rich in features; it is a stable structure which does not change with the age (8 to 70). It doesn't change its shape with facial expressions. Furthermore, the ear is larger compared to fingerprints and can be easily captured [11].

In recent years, some approaches have been developed for ear recognition. Burge and Burger [5] proposed an adjacency graph, which is built from the Voronoi diagram of the ear's edge segments, to describe the ear. Ear recognition is done by subgraph matching. Hurley et al. [12] applied force field transform to ear images and wells and channels are shown to be invariant to affine transformations. Victor et al. [18] used Principal Component Analysis to ear images. All of these papers have used 2D intensity images and, therefore, the performance of these systems is greatly affected by imaging conditions. However ears can be imaged in 3-D from a distance and we can develop a robust ear biometric.

In this paper, we introduce a new local surface descriptor for 3D ear representation. We calculate the local surface descriptors only for the feature points which are defined as the local minimum and maximum of shape indexes calculated from principal curvatures [8]. Our approach starts from extracting feature points from range images, then define the local surface patch as the feature point and its neighbors, next calculate local surface properties which are 2D histogram, surface type and the centroid. The 2D histogram consists of shape indexes and angles between the normal of reference point and that of its neighbors. By comparing local surface patches, we find the potential corresponding local surface patches. Finally, we estimate the transformation based on the corresponding surface patches and calculate the match quality between the hypothesized model and test image.

The rest of the paper is organized as follows. We introduce the related work and motivation in Section 2. In Section 3, our approach to represent the free-form surfaces and matching the surface patches is presented. Section 4 gives the experiment results to demonstrate the effectiveness of our approach. Conclusion is provided in Section 5.

2. Related work and motivation

2.1 Related work in 3D object recognition

In 3D object recognition, the key problems are how to represent free-form surfaces effectively and how to match the surfaces using the selected representation. In the early years of 3D computer vision, the representation schemes included Wire-Frame, Constructive Solid Geometry (CSG), Extended Gaussian Image (EGI), Generalized Cylinders, planar faces [2] and Superquadric [17] [3]. All of these are not suitable for representing free-from surfaces. The ear can be thought of as a rigid free-form object.

Stein and Medioni [16] used two different types of primitives: 3-D curves and splashes, for representation and matching. 3-D curves are edges corresponding to the depth and orientation discontinuities. For smooth areas, splash is defined as surface normals along contours of different radii. Both of them can be encoded by a set of 3D super segments, which are described by the curvature and torsion angles of a super segment. The 3D super segments are indexed into a hash table for fast retrieval and matching. Therefore, all the model information is recorded in the hash table. Hypothesis is generated by casting votes to the hash table and bad hypotheses are removed by estimating rigid transformation. Chua and Jarvis [6] used point signature, which can describe the structural neighborhood of a point, to represent 3D free-form objects. Point signature is one-dimensional signed distance profile with respect to the rotation angle defined by the angle between the normal vector and the reference vector. Recognition is performed by matching the signatures of points on the scene surfaces to those of points on the model surfaces. The maximum and minimum values of the signatures are used as indexes to a 2D table for fast retrieval and matching.

Johnson and Hebert [14] presented the spin image (SI) which is really a 2D histogram. Given an oriented point on the surface, its shape is described by two parameters. One is the distance to the tangent plane of the oriented point from its neighbors; the other is the distance to the normal vector of the oriented point. There are three steps: spin image generation, correspondence points finding and verification. First, spin images are calculated at every vertex of the model surfaces. Then the corresponding point pair is found by computing the correlation coefficient of two spin images centered at those two points. Next the corresponding pairs are filtered by using geometric constraints. Finally, a rigid transformation is computed and a modified Iterative Closest Point (ICP) algorithm is used for verification. In order to speed up the matching process, principal component analysis (PCA) is used to compress spin images. Salvador et al. [7] proposed the spherical spin image (SSI) which maps the spin image to points onto a unit sphere. Corresponding points can be found by computing the angle between two SSI. Yamany et al. [19] introduced the surface signature which is also a 2D histogram. One parameter is the distance between the center point and every surface point. The other one is the angle between the normal of the center point and every surface point. Signature matching can be done by template matching. Zhang and Hebert [20] introduced harmonic shape images (HSI) which are 2D representation of 3D surface patches. HSI are unique and they can preserve the shape and continuity of the underlying surfaces. Surface matching can be simplified to matching harmonic shape images.

2.2 Motivation

Some of above approaches generated surface signatures for very point on the surface, which is computationally expensive. Moreover, the surface signatures are similar if the two points are close to each other in the 3D space. Therefore, it's not necessary to calculate surface signature at every vertex on the surface. In our approach, we only calculate surface signatures at feature points which are defined as the local minimum and maximum of shape indexes. Because the imaged surface is only a sampling of the actual surface, it's almost impossible for a point in an image at a certain viewpoint to appear in another image at a different viewpoint (even the same viewpoint). It's more reasonable to find corresponding surfaces. Therefore, we use local surface patches as our representation.

Motivated by [16] [14], our approach has two steps: off-line preprocessing and on-line recognition. During the first step, we extract feature points, calculate some features for every local surface patch and save them into the model database. During the online recognition, we find potential corresponding surface patches by comparing the test local surface patches with model local surface patches. In our approach, we use shape indexes, normal angles as our basic features to represent local surface properties, since shape indexes and angles between surface normals are invariant to rigid transformation.

Furthermore, when we calculate the rigid transformation, we use the centroid of the local surface patch instead of using the 3D coordinates of the feature point. The centroid is less sensitive to the noise since it is the average of 3D coordinates of the local surface patch.

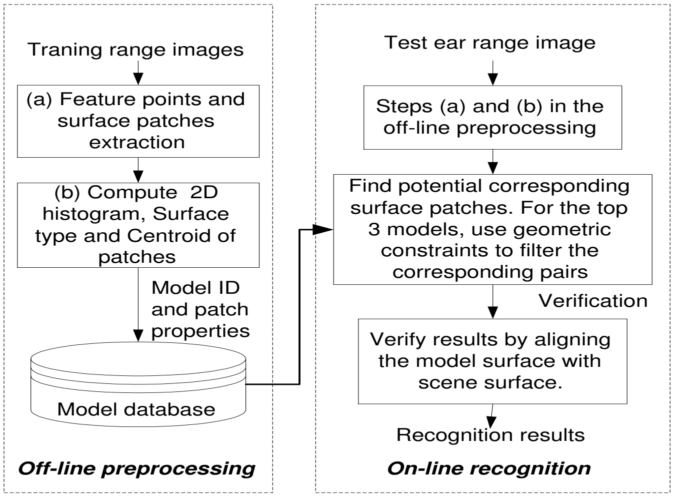

Traning range images
(a) Feature points and surface patches extraction
(b) Compute 2D histogram, Surface type and Centroid of patches
Model ID and patch properties
Model database
Off-line preprocessing
Test ear range image
Steps (a) and (b) in the off-line preprocessing
Find potential corresponding surface patches. For the top 3 models, use geometric constraints to filter the corresponding pairs
Verification
Verify results by aligning the model surface with scene surface.
Recognition results
On-line recognition

Figure 1. System diagram.

3. Technical approach

Our approach has two phases: off-line preprocessing and on-line recognition. The block diagram is illustrated in Figure 1.

3.1 Feature points extraction

In our approach, feature points are defined as local minimum and maximum of shape indexes, which can be calculated from principal curvatures [8]. In order to estimate curvatures, we fit a biquadratic surface (1) to a local window and use the least square method to estimate the parameters of the quadratic surface, and then use differential geometry to calculate the surface normal, Gaussian and mean curvatures and principal curvatures [9] [17]. Based on differential geometry, surface normal $\vec{n}$, Gaussian curvature K, mean curvature H, principal curvatures $k_{1,2}$ are given by (2), (3), (4) and (5) respectively.

$$f(x,y) = ax^2 + by^2 + cxy + dx + ey + f \quad (1)$$

$$\vec{n} = \frac{(-f_x, -f_y, 1)}{\sqrt{1 + f_x^2 + f_y^2}} \quad (2)$$

$$K = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1 + f_x^2 + f_y^2)^2} \quad (3)$$

$$H = \frac{f_{xx} + f_{yy} + f_{xx} f_y^2 + f_{yy} f_x^2 - 2 f_x f_y f_{xy}}{2(1 + f_x^2 + f_y^2)^{1.5}} \quad (4)$$

$$k_{1,2} = H \pm \sqrt{H^2 - K} \quad (5)$$

Shape index (S_i) at a point p is defined by (6) where k_1 and k_2 are maximum and minimum principal curvatures respectively.

$$S_i(p) = \frac{1}{2} - \frac{1}{\pi} tan^{-1} \frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \quad (6)$$

Within a 5 × 5 window, the center point is marked as a feature point if its shape index is higher or lower than those of its neighbors.

The feature points extraction results are shown in Figure 2 where the feature points are marked by red plus sign. In order to see the feature points' location, we enlarge the two images. From the Figure 2, we can clearly see that some feature points corresponding to the same physical area appear in both images.

3.2 Local surface patches

We define a "local surface patch" as the region consisting of a feature point P and its neighbors N. A local surface patch is shown in Figure 3. The neighbors should satisfy these two conditions,

$$\begin{aligned} N &= \{pixels \quad N, ||N - P|| \leq \epsilon_1\} \\ & and \ acos(n_p \bullet n_n < A), \end{aligned} \quad (7)$$

where n_p and n_n are the surface normal vectors at point P and N. The two parameters ϵ_1 and A are important since they determine how the local surface patch is resistant to clutter and occlusion. Johnson [14] discussed the choices for the two parameters. For every local surface patch, we compute the shape indexes and normal angles between point P and its neighbors. Then we can form a 2D histogram. One axis of this histogram is the shape index which is in the range [0,1]; the other is the dot product of surface normal vectors at P and N which is in the range [-1,1]. In order to reduce the effect of the noise, we use bilinear interpolation when we calculate the 2D histogram [14].

We also compute the centroid of local surface patches. For the feature point, we can get the surface type based on the Gaussian and mean curvatures [1] [4]. There are 8 surface types determined by the signs of Gaussian and mean curvatures given in Table 1. Note that a feature point and the centroid of a patch may not coincide.

In summary, every local surface patch is described by a 2D histogram, surface type and the centroid. The local surface patch encodes the geometric information of a local surface.

3.3 Off-line model building

Considering the uncertainty of location of a feature point, we repeat the above process to calculate de-

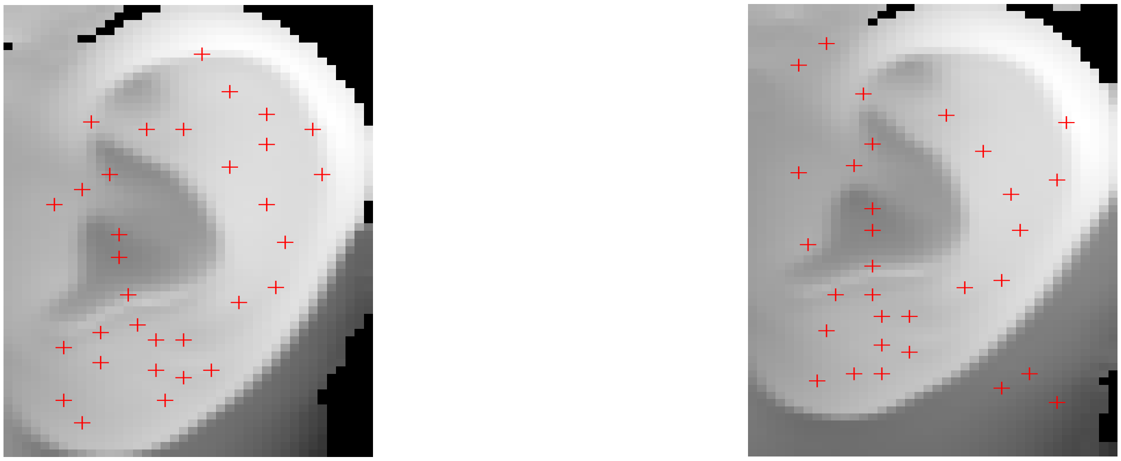


Figure 2. Feature points location in two range images of the same ear shown as gray scale images. The darker points are away from the camera and the lighter ones are closer.

Table 1. Surface type Tp based on the signs of Mean curvature(H) and Gaussian curvature(K).

Mean Curvature H	Gaussian Curvature K		
	$K > 0$	$K = 0$	$K < 0$
$H < 0$	Peak Tp=1	Ridge Tp=2	Saddle Ridge Tp=3
$H = 0$	None Tp=4	Flat Tp=5	Minimal Tp=6
$H > 0$	Pit Tp=7	Valley Tp=8	Saddle Valley Tp=9

scriptor of local surface patches for neighbors of feature point P and save these descriptions into the model database. For each model object, we repeat the same process to build the model database.

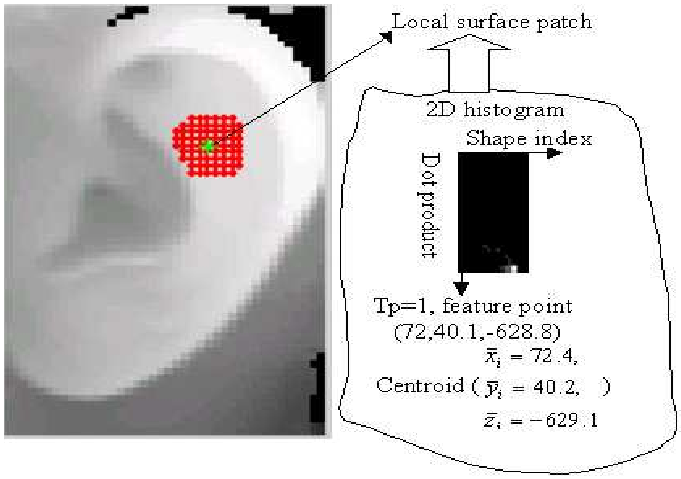


Figure 3. Illustration of Local Surface Patch.

3.4 Recognition

Given a test range image, we repeat the above steps and get local surface patches. Considering the inaccuracy of feature points' location, we also extract local surface patches from neighbors of feature points. Then we compare them with all of the local surface patches saved in the model database. This comparison is based on the surface type and histogram dissimilarity. Since histogram can be thought of as an approximation of probability distributed function, we use statistical method to assess the dissimilarity between two probability distributions. The $\chi^2 - divergence$ is among the most prominent divergence used in statistics to assess the dissimilarity between two probability density functions. We use it to measure the dissimilarity between two observed histograms Q and V, which is defined by (8) [15].

$$\chi^2(Q,V) = \sum_i \frac{(q_i - v_i)^2}{q_i + v_i} \tag{8}$$

From (8), we know the dissimilarity is between 0 and 2. If the two histograms are exactly same, the dissimilarity will be zero. If the two histograms don't overlap with each other, it will achieve the maximum value 2.

For every local surface path from the test ear, we choose the local surface patch from the database with minimum dissimilarity and same surface type as the possible corresponding patch. Using the above steps, we get the number of possible corresponding local surface patches for each model. For the top 3 models which get three highest numbers, we filter the possible corresponding pairs based on the geometric constraints given below.

$$d_{C_1,C_2} = |d_{S_1,S_2} - d_{M_1,M_2}| < \epsilon_2, \tag{9}$$

Where d_{S_1,S_2} and d_{M_1,M_2} are Euclidean distance between centroids of two surface patches. For two correspondences $C_1 = \{S_1, M_1\}$ and $C_2 = \{S_2, M_2\}$ where S means test surface patch and M means model surface patch, they should satisfy (9) if they are consistent corresponding pairs. Thus, we use geometric constraints to partition the potential corresponding pairs into different groups. The largest group would be more likely to be the true corresponding pair.

Given a list of corresponding pairs $L = \{C_1, C_2, \ldots, C_n\}$, the grouping procedure for every pair in the list is as follows: Initialize each pair of a group. For every group, add other pairs to it if they satisfy (9). Repeat the same procedure for every group. Select the group which has the largest size.

3.5 Verification

After filtering the corresponding pairs, we get the largest group which is at least three potential matched pairs of local surface patches. By using quaternion representation [10], we calculate the rotation matrix and translation vector. Applying this transformation to the model object, we get a transformed data set. For every point in this dataset, we search the closest point in the test image. If the Euclidean distance between them is less than ϵ_3, they are considered as corresponding points. Thus, we can get the match quality, MQ defined below.

$$MQ = \frac{\#\ of\ corresponding\ points}{\#\ of\ total\ model\ points} \tag{10}$$

In order to speed up the nearest point search process, we use k-d tree.

4. Experiment results

4.1 Data and parameters

We use real range data acquired using a range finder. There are ten ears in our database and they are E0, E1, E2, E3, E4, E5, E6, E7, E8, E9 where the number represents Model ID. The model ears are shown in Figure 5, and the test ears are shown in Figure 6. For Figures 5 and 6, we only use z coordinates to show the ears. The 3D surfaces of the model ears are shown in Figure 4.

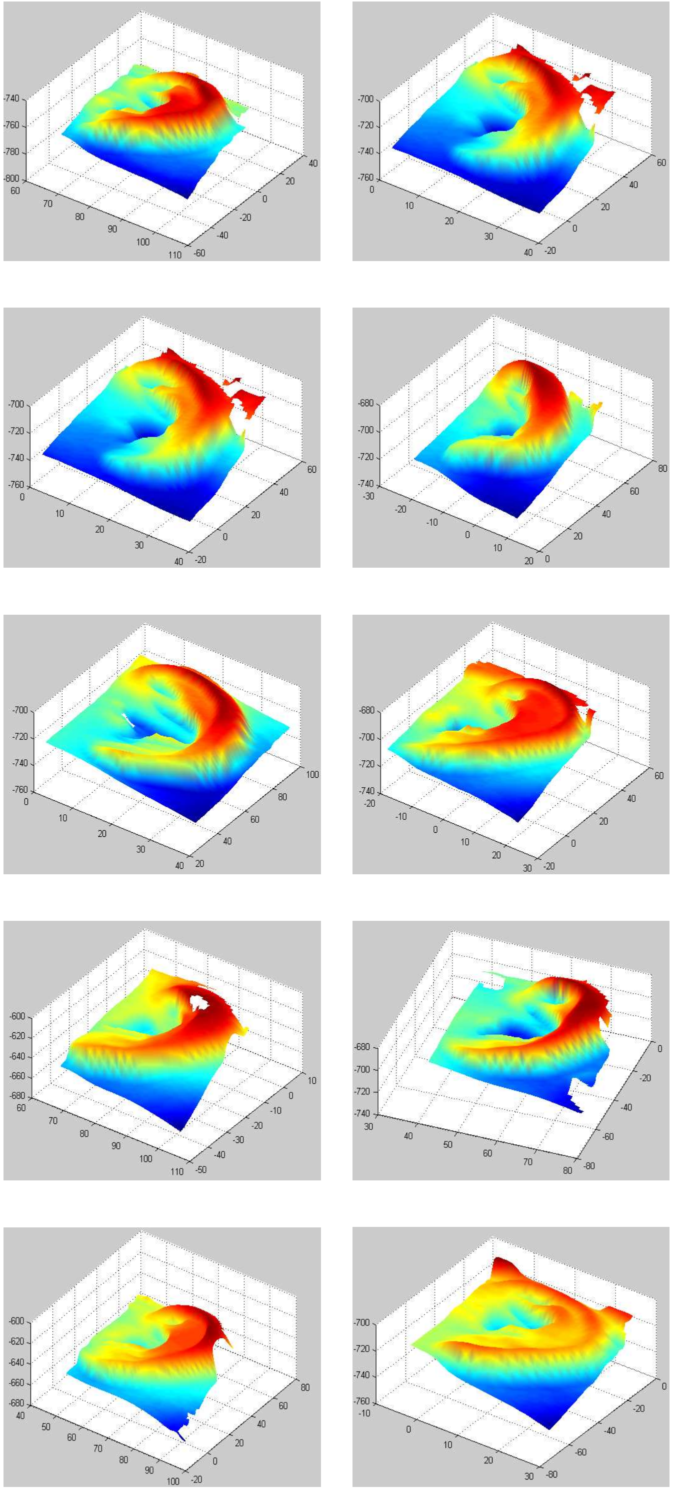


Figure 4. 3D surfaces of model ears E0-E9 (from left to right and top to bottom).

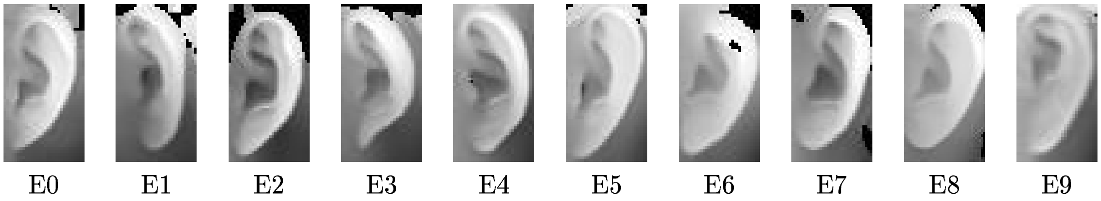


Figure 5. Model ear range images E0-E9.

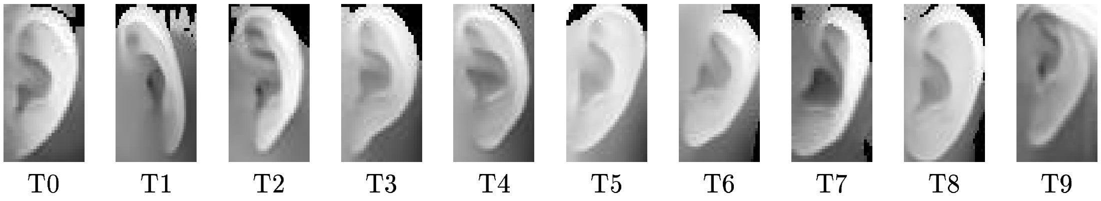


Figure 6. Test ear range images T0-T9 corresponding to E0-E9.

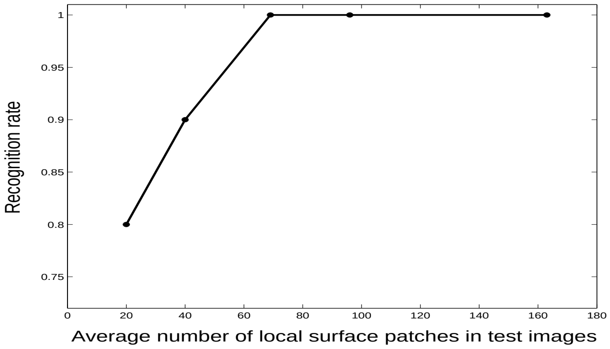


Figure 7. Recognition rate vs. the average number of local surface patches in test images.

The parameters of our approach are $\epsilon_1 = 5.8mm$, $A = \pi/3$, $\epsilon_2 = 9.4mm$ and $\epsilon_3 = 4.8mm$. The bin size of the two dimensions of 2D histogram is 0.06. The average size of local surface patch is 47 pixels.

4.2 Results

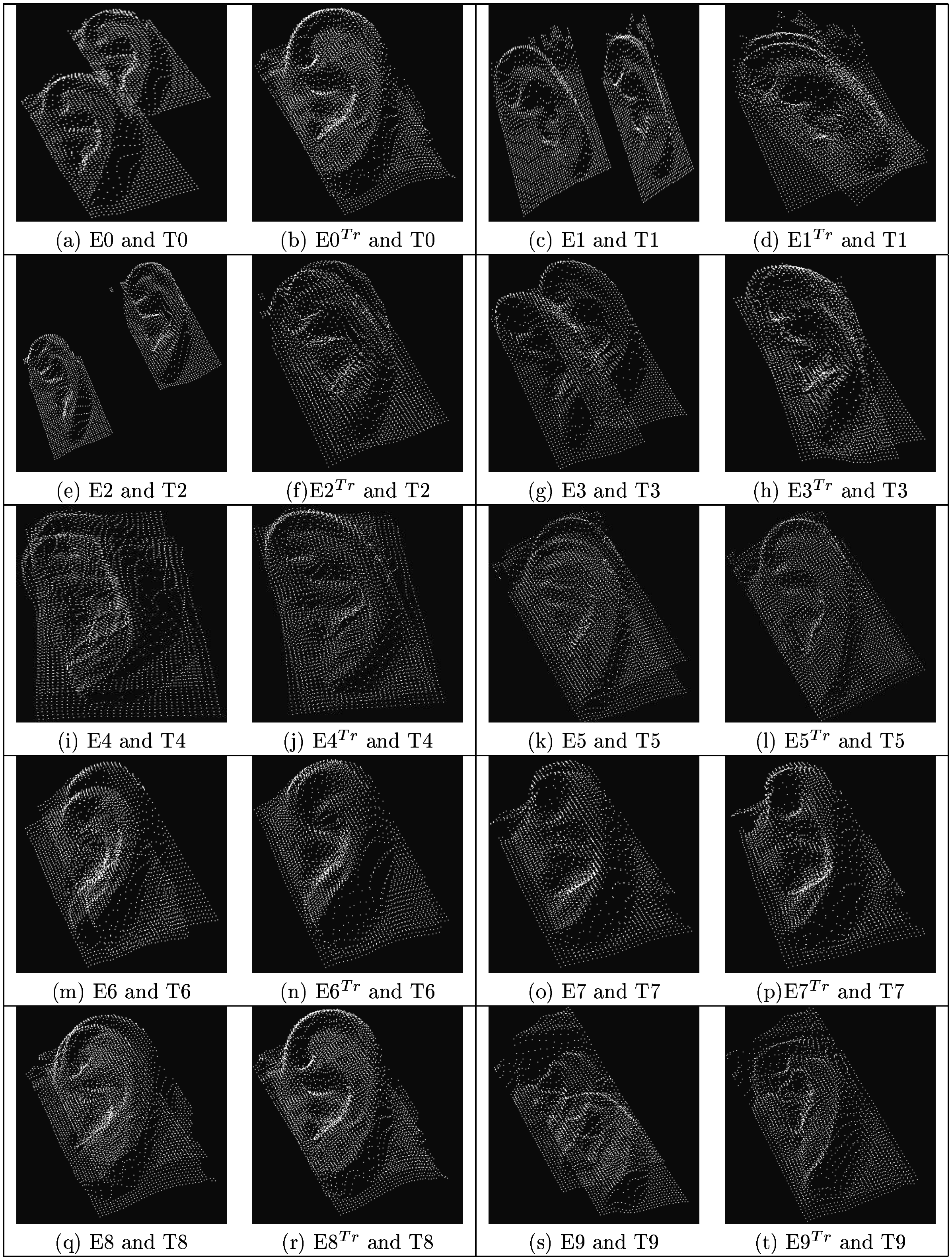
The recognition results on real data are shown in Table 2. In Table 2, the number in the parenthesis means the number of local surface patches. From this table, we can clearly see that most of the highest number of corresponding pairs go to the right ear models. By estimating the rigid transformation, we calculate the match quality listed in Table 2. We choose the model with the maximum match quality as the recognized ear. It's clearly seen that we achieve 100% recognition rate for our dataset.

As mentioned in Section 3.4 to allow for uncertainty in location of feature points, we extract local surface patches (LSP) from neighbors of feature points. Choosing different neighborhood size, we repeat the experiments and get the relationship between recognition rate and the average number of LSP in test images. The result is shown in Figure 7. From Figure 7, we can see better recognition results are obtained with a larger number of LSPs.

We show the visualization of our recognition results in Figure 8. In order to evaluate our results, we display the model ear and test ear in the same image, the transformed model and test ear in the same image. With our programs, we can view them at different viewpoints. In Figure 8, we only display them at a certain viewpoint. For the ear E0, the model ear and test ear are shown in Figure 8(a); the transformed model ear and test ear are shown in Figure 8(b). We clearly see that the transformed model ear is well aligned with the test ear. For the ear E1, the model ear and test ear are shown in Figure 8(c); the transformed model ear and test ear are shown in Figure 8(d). We can see that E0 is a better fit to the test ear T0 than E1 to T1. Similar results are shown in Figure 8(e) and (f) for the ear E2, in Figure 8(g) and (h) for the ear E3, in Figure 8(i) and (j) for the ear E4, in Figure 8(k) and (l) for the ear E5, in Figure 8(m) and (n) for the ear E6, in Figure 8(o) and (p) for the ear E7, in Figure 8(q) and (r) for the ear E8, in Figure 8(s) and (t) for the ear E9.

5. Conclusions

We have presented an approach for recognition of 3D ears. We have used a new integrated local surface patch representation. Through experiments, we see that the local surface patch is a good local surface descriptor, since we can get good corresponding pairs



(a) E0 and T0

(b) E0^{Tr} and T0

(c) E1 and T1

(d) E1^{Tr} and T1

(e) E2 and T2

(f)E2^{Tr} and T2

(g) E3 and T3

(h) E3^{Tr} and T3

(i) E4 and T4

(j) E4^{Tr} and T4

(k) E5 and T5

(l) E5^{Tr} and T5

(m) E6 and T6

(n) E6^{Tr} and T6

(o) E7 and T7

(p)E7^{Tr} and T7

(q) E8 and T8

(r) E8^{Tr} and T8

(s) E9 and T9

(t) E9^{Tr} and T9

Figure 8. Visualization of recognition results.^{Tr} means transformed model ears.

Table 2. Recognition results.

Test ears	Results (Top 3 matches)			
T0(50)	Model ID	0	8	5
	No. of Pairs	26	7	6
	Match Quality	100%	66%	81%
T1(162)	Model ID	1	8	5
	No. of Pairs	31	29	27
	Match Quality	97%	46%	35%
T2(70)	Model ID	2	4	3
	No. of Pairs	15	12	11
	Match Quality	92%	44%	43%
T3(63)	Model ID	8	5	3
	No. of Pairs	17	10	7
	Match Quality	53%	35%	91%
T4(182)	Model ID	4	8	3
	No. of Pairs	28	11	9
	Match Quality	99%	80%	
T5(120)	Model ID	5	8	4
	No. of Pairs	58	15	14
	Match Quality	99%	38%	46%
T6(103)	Model ID	6	8	4
	No. of Pairs	30	18	13
	Match Quality	99%	31%	26%
T7(54)	Model ID	7	8	2
	No. of Pairs	29	9	3
	Match Quality	97%	63%	
T8(153)	Model ID	8	5	6
	No. of Pairs	97	14	10
	Match Quality	97%	90%	
T9(109)	Model ID	9	8	5
	No. of Pairs	27	23	23
	Match Quality	94%	35%	49%

based on comparing local surface patches. The experimental results show the potential of our approach for robust ear recognition in 3D.

References

[1] P. Besl and R. Jain. Segmentation through variable-order surface fitting. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(2):167–192, 1988.

[2] B. Bhanu. Representation and shape matching of 3-D objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(3):340–351, 1984.

[3] B. Bhanu and C. Ho. CAD-based 3d object representation for robot vision. *IEEE Computer*, 20(8):19–35, 1987.

[4] B. Bhanu and L. Nuttall. Recognition of 3-D objects in range images using a butterfly multiprocessor. *Pattern Recognition*, 22(1):49–64, 1989.

[5] M. Burge and W. Burger. Ear biometrics in computer vision. *Proc. Int. Conf. on Pattern Recognition*, 2:822–826, 2000.

[6] C. Chua and R. Jarvis. Point signatures: A new representation for 3D object recognition. *International Journal of Computer Vision*, 25(1):63–85, 1994.

[7] S. Correa and L. Shapiro. A new signature-based method for efficient 3-D object recognition. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1:769–776, 2001.

[8] C. Dorai and A. Jain. COSMOS-a representation scheme for free-form surfaces. *Proc. Int. Conf. on Computer Vision*, pages 1024–1029, 1995.

[9] P. Flynn and A. Jain. On reliable curvature estimation. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 110–116, 1989.

[10] B. Horn. Close-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4(4):629–642, 1987.

[11] D. Hurley. *Force Field Feature Extraction for Ear Biometrics.* PhD thesis. Dept. of Electronics and Computer Science,Univ. of Southampton,UK, 2001.

[12] D. Hurley, M. Nixon, and J. Carter. Automatic ear recognition by force field transformations. *IEE Colloquium on Visual Biometrics*, pages 7/1 –7/5, 2000.

[13] A. Iannarelli. *Ear Identification.* Forensic Identification Series. Paramont Publishing Company, 1989.

[14] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(5):433–449, 1999.

[15] B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[16] F. Stein and G. Medioni. Structural indexing: efficient 3-D object recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(2):125–145, 1992.

[17] M. Suk and S. Bhandarker. *Three-Dimensional object recognition from range images.* Springer-Verlag, 1992.

[18] B. Victor, K. Bowyer, and S. Sarkar. An evaluation of face and ear biometrics. *Proc. Int. Conf. on Pattern Recognition*, 1:429–432, 2002.

[19] S. M. Yamany and A. Farag. Free-form surface registration using surface signatures. *Proc. Int. Conf. on Computer Vision*, 2:1098–1104, 1999.

[20] D. Zhang and M. Herbert. Harmonic maps and their applications in surface matching. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2:524–530, 1999.

Multimodal Biometric Authentication Methods: A COTS Approach

M. Indovina[1], U. Uludag[2], R. Snelick[1], A. Mink[1], A. Jain[2]
[1]National Institute of Standards and Technology, [2]Michigan State University
{mindovina, rsnelick, amink}@nist.gov, {uludagum, jain}@cse.msu.edu

Abstract

We examine the performance of multimodal biometric authentication systems using state-of-the-art Commercial Off-the-Shelf (COTS) fingerprint and face biometrics on a population approaching 1000 individuals. Prior studies of multimodal biometrics have been limited to relatively low accuracy non-COTS systems and populations approximately 10% of this size. Our work is the first to demonstrate that multimodal fingerprint and face biometric systems can achieve significant accuracy gains over either biometric alone, even when using already highly accurate COTS systems on a relatively large-scale population. In addition to examining well-known multimodal methods, we introduce novel methods of fusion and normalization that improve accuracy still further through population analysis.

1. Introduction

It has recently been reported [1] to the U.S. Congress that approximately two percent of the population does not have a legible fingerprint and therefore cannot be enrolled into a fingerprint biometrics system. The report recommends a system employing dual biometrics in a layered approach. Use of multiple biometric indicators for identifying individuals, so-called multimodal biometrics, has been shown to increase accuracy [2, 3, 4], and would decrease vulnerability to spoofing while increasing population coverage.

The key to multimodal biometrics is the fusion (i.e., combination) of the various biometric mode data at the feature extraction, match score, or decision level [4]. Feature level fusion combines feature vectors at the representation level to provide higher dimensional data points when producing the match score. Match score level fusion combines the individual scores from multiple matchers. Decision level fusion combines accept or reject decisions of individual systems.

Our methodology for testing multimodal biometric systems focuses on the match score level [2]. This approach has the advantage of utilizing as much information as possible from each single-mode biometric, while at the same time enabling the integration of proprietary COTS systems.

Published studies examining fusion techniques have been limited to small populations (~100 individuals), while employing low performance non-commercial biometric systems. In this paper we investigate the performance gains achievable by COTS-based multimodal biometric systems using a relatively large (~1000 individuals) population. Section two and three describe the traditional and novel normalization and fusion methods that we employed for match score combination. New methods for *adaptive normalization* and fusion using user-level weighting based on the *wolf-lamb* [5] concept are introduced and compared. In section four we provide a performance analysis of these multimodal methods and investigate performance variability attributable to population differences.

2. Normalization

A normalization step is generally necessary before the raw scores originating from different matchers can be combined in the fusion stage. For example, if one matcher yields scores in the range [100, 1000] and another matcher in the range [0, 1], fusing the scores without any normalization effectively eliminates the contribution of the second matcher. We present three well-known normalization methods, and a 4th novel method, which we call *adaptive normalization* that uses the genuine and impostor distributions.

We denote a raw matcher score as s from the set S of all scores for that matcher, and the corresponding normalized score as n. Different sets are used for different matchers. The abbreviations (such as MM) next to the normalization method names are used throughout the remainder of this paper.

Min-Max (MM). This method maps the raw scores to the [0, 1] range. max(S) and min(S) specify the end points of the score range (vendors generally provide these values):

$$n = \frac{s - min(S)}{max(S) - min(S)}$$

Z-score (ZS). This method transforms the scores to a distribution with mean of 0 and standard deviation of 1. $mean()$ and $std()$ denote the mean and standard deviation operators:

$$n = \frac{s - mean(S)}{std(S)}$$

Tanh (TH). This method is among the so-called *robust* statistical techniques [6]. It maps the scores to the (0, 1) range:

$$n = \frac{1}{2}\left[tanh\left(0.01\frac{(s - mean(S))}{std(S)} \right) + 1 \right]$$

Adaptive (AD). The errors of individual biometric matchers stem from the overlap of the genuine and impostor distributions as shown in Fig. 1. This region is characterized with its center c and its width w. To decrease the effect of this overlap on the fusion algorithm, we propose to use an adaptive normalization procedure that aims to increase the separation of the genuine and impostor distributions, as indicated by the block arrows in Fig. 1., while still mapping scores to [0,1].

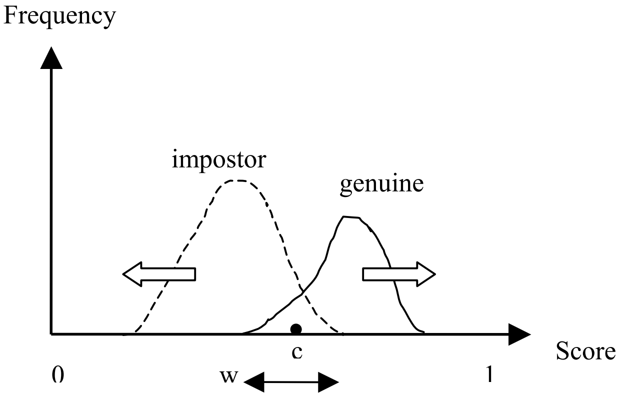


Fig. 1. Overlap of genuine and impostor distributions.

This adaptive normalization is formulated as

$$n_{AD} = f(n_{MM})$$

where $f()$ denotes the mapping function which is used on the MM normalized scores. We have considered the following three functions for $f()$. These functions use two parameters of the overlapped region, c and w, which can be provided by the vendors or estimated by the integrator from data sets appropriate for the specific application. In this work, we act as the integrator.

Two-Quadrics (QQ). This function is composed of 2 quadratic segments that change concavity at c (Fig. 2).

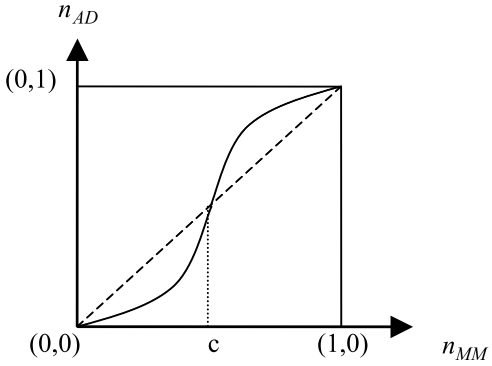


Fig. 2. Mapping function for QQ.

$$n_{AD} = \begin{cases} \frac{1}{c} n_{MM}^2 , & n_{MM} \leq c \\ c + \sqrt{(1-c)(n_{MM} - c)}, & \text{otherwise} \end{cases}$$

For comparison, note that the identity function, $n_{AD} = n_{MM}$, is shown by the dashed line.

Logistic (LG). Here, $f()$ takes the form of a logistic function. The general shape of the curve is similar to that shown for function QQ in Fig. 2. It is formulated as

$$n_{AD} = \frac{1}{1 + A \cdot e^{-B \cdot n_{MM}}}$$

where the constants A and B are calculated as

$$A = \frac{1}{\Delta} - 1 \text{ and } B = \frac{\ln A}{c}$$

Here, $f(0)$ is equal to the constant Δ, which is selected to be a small value (0.01 in this study). Note the inflection point of the logistic function occurs at c, the center of the overlapped region.

Quadric-Line-Quadric (QLQ). The overlapped zone, w, is left unchanged while the other regions are mapped with two quadratic function segments (Fig. 3):

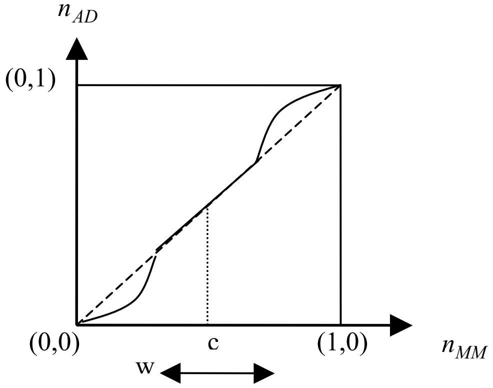


Fig. 3. Mapping function for QLQ.

$$n_{AD} = \begin{cases} \dfrac{1}{(c-\frac{w}{2})} n_{MM}^2, & n_{MM} \le (c-\frac{w}{2}) \\ n_{MM}, & (c-\frac{w}{2}) < n_{MM} \le (c+\frac{w}{2}) \\ (c+\frac{w}{2}) + \sqrt{(1-c-\frac{w}{2})(n_{MM}-c-\frac{w}{2})}, & o/w \end{cases}$$

3. Fusion

We experimented with the five different fusion methods summarized below. The first three are well-known fusion methods; the last two are novel and they utilize the performance of individual matchers in weighting their contributions. As we progress from the first three methods to the fifth, the amount of data necessary to apply the fusion method increases.

Our notation is as follows: n_i^m represents the normalized score for the matcher m ($m = 1, 2, ..., M$, where M is the number of different matchers) and for the user i ($i = 1, 2, ..., I$, where I is the number of individuals in the database). The fused score is denoted as f_i.

Simple Sum (SS). Scores for an individual are summed:

$$f_i = \sum_{m=1}^{M} n_i^m, \forall i$$

Min Score (MIS). Choose the minimum of an individual's scores:

$$f_i = min(n_i^1, n_i^2, ..., n_i^M), \forall i$$

Max Score (MAS). Choose the maximum of an individual's scores:

$$f_i = max(n_i^1, n_i^2, ..., n_i^M), \forall i$$

Matcher Weighting (MW). Matcher weighting-based fusion makes use of the Equal Error Rate (EER). Denote the EER of matcher m as e^m, $m = 1, 2, ..., M$ and the weight w^m associated with a matcher m is calculated as

$$w^m = \frac{\dfrac{1}{\sum_{m=1}^{M} \frac{1}{e^m}}}{e^m} \qquad (1)$$

Note that $0 \le w^m \le 1, \forall m$, $\sum_{m=1}^{M} w^m = 1$ and the weights are inversely proportional to the corresponding errors; the weights for *more accurate* matchers are higher than those of *less accurate matchers* (Although the EER value alone may not be a good estimator for the accuracy of a matcher, we chose to use it for spanning the amount of data available to the integrator mentioned above). The MW fused score is calculated as

$$f_i = \sum_{m=1}^{M} w^m n_i^m, \forall i$$

User Weighting (UW). The User Weighting fusion method applies weights to individual matchers differently for every user (individual). Previously, Ross and Jain [7] proposed a similar scheme, but they *exhaustively* search a coarse sampling of the weight space, where weights are multiples of 0.1. Their method can be prohibitively expensive if the number of fused matchers, M, is high, since the weight space is $\Re^M$; further, coarse sampling may hinder the calculation of an optimal weight set. In our method, the UW fused score is calculated as

$$f_i = \sum_{m=1}^{M} w_i^m n_i^m, \forall i$$

where w_i^m represents the weight of matcher m for user i.

The calculation of these user-dependent weights make use of the *wolf-lamb* concept introduced by Doddington, et al. [5] for unimodal speech biometrics. They label the users who can be imitated easily as *lambs*; *wolves* on the other hand are those who can successfully imitate some others. Lambs and wolves decrease the performance of biometric systems since they lead to false accepts.

We extend these notions to multimodal biometrics by developing a metric of *lambness* for every user and matcher, (i,m), pair. This lambness metric is then used to calculate weights for fusion. Thus, if user i is a *lamb* (can be imitated easily by some *wolves)* in the space of matcher m, the weight associated with this matcher is decreased. The main aim is to decrease the lambness of user i in the space of combined matchers.

We assume that for every (i,m) pair, the mean and standard deviation of the associated genuine and impostor distributions are known (or can be calculated, as is done in this study). Denote the means of these distributions as ${}^{gen}\mu_i^m$ and ${}^{imp}\mu_i^m$, respectively, and denote the standard deviations as ${}^{gen}\sigma_i^m$ and ${}^{imp}\sigma_i^m$, respectively.

We use the d-prime metric [8] as a measure of the separation of these two distributions in formulating the lambness metric as:

$$d_i^m = \frac{{}^{gen}\mu_i^m - {}^{imp}\mu_i^m}{\sqrt{({}^{gen}\sigma_i^m)^2 + ({}^{imp}\sigma_i^m)^2}}$$

If d_i^m is small, user i is a lamb for some wolves; if d_i^m is large, i is not a lamb. We structure the user weights to be proportional to this lambness metric as follows

$$w_i^m = \frac{1}{\sum_{m=1}^{M} d_i^m} \cdot d_i^m \qquad (2)$$

Note that $0 \le w_i^m \le 1, \forall i, \forall m$, and $\sum_{m=1}^{M} w_i^m = 1, \forall i$.

Fig. 4 shows the location of potential wolves for a specific (i,m) pair with a block arrow, along with the associated genuine and impostor distributions. This user dependent weighting scheme addresses the issue of matcher-user relationship: namely, a user can be lamb for a specific matcher, but also can be a wolf for some other matcher. We find the user weights by measuring the respective threat of wolves *living* in different matcher spaces for every user.

4. Experimental Results

4.1. Databases

Our experiments were conducted on a population of consistently paired fingerprint and facial images from two groups of 972 individuals, using our previously developed test methodology and framework [2]. Since the paired fingerprint and facial images come from different individuals, we are assuming that they are statistically independent – a widely accepted practice. The images were taken from two separate groups of 972 individuals, with the first group contributing a pair of facial images and the second a pair of fingerprint images. This creates a database of 972 *virtual* individuals. Each pair consists of a primary and a secondary image, with all primary images assigned to the *target* set, and all secondary images assigned to the *query* set.

Match scores were generated from four COTS biometric systems – three fingerprint and one face. For each biometric system, all query set images were matched against all target set images, yielding 972 genuine scores (correct matches) and 943,812 imposter scores.

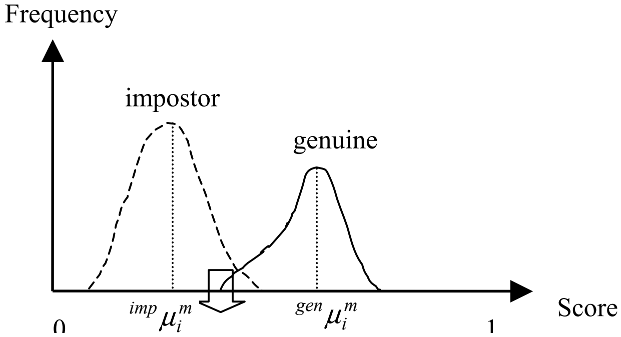


Fig. 4. Distributions for a (user, matcher) pair: the arrow indicates location of wolves for lamb i

4.2. Approach

Among the three adaptive normalization methods (QQ, QLQ and LG), the QLQ method gave the best results in our experiments, so it is selected as the representative method.

We carried out all possible permutations of (normalization, fusion) techniques on our database of 972 users. Table 1 shows the EER values for these permutations. Note that EER values for the 3 individual fingerprint matchers and the face matcher are found to be 3.96%, 3.72%, 2.16% and 3.76%, respectively. The best EER values in individual columns are indicated with **bold** typeface; the best EER values in individual rows are indicated with a star (*) symbol.

Table 1. EER values for permutations (%).

Normalization Technique	*Fusion Technique*				
	SS	MIS	MAS	MW	UW
MM	0.99	5.43	0.86	**1.16**	***0.63**
ZS	*1.71	5.28	1.79	1.72	1.86
TH	1.73	**4.65**	1.82	*1.50	1.62
QLQ	**0.94**	5.43	***0.63**	**1.16**	***0.63**

4.3. Normalization

Figures 5-9 show the effect of each normalization method on system performance for different (but fixed) fusion methods. The ROCs (Receiver Operating Characteristics) for the individual fingerprint matchers and the face matcher are also shown for better comparison.

For UW fusion (Fig. 9), the scatter plot of user weights (Fig. 10) form a distinctive band-like behavior for each fingerprint matcher V1, V2, V3, and the face matcher. The mean user weights for the individual biometric matchers, calculated from (2), are 0.14, 0.64, 0.17 and 0.05, respectively, which implies that on average, fingerprint matcher V2 is the safest for the lambs; whereas the space of the face matcher is filled with wolves (i.e., those waiting to be falsely accepted as one of the lambs). Note that individual matcher performance, shown in the previous ROC curves, is not reflected directly in the set of user weights and their means. Namely, V2 has a higher mean user weight than V3, despite V3's generally better ROC.

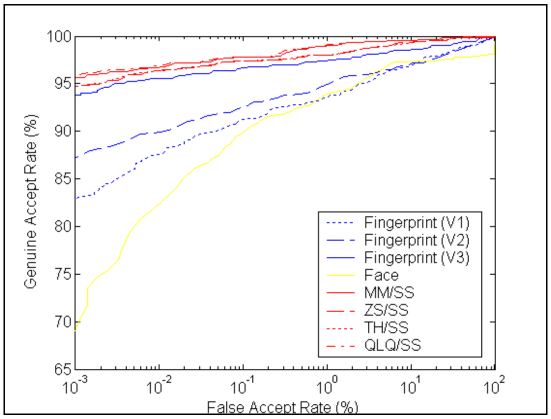


Fig. 5. ROC curves for SS, normalization varied.

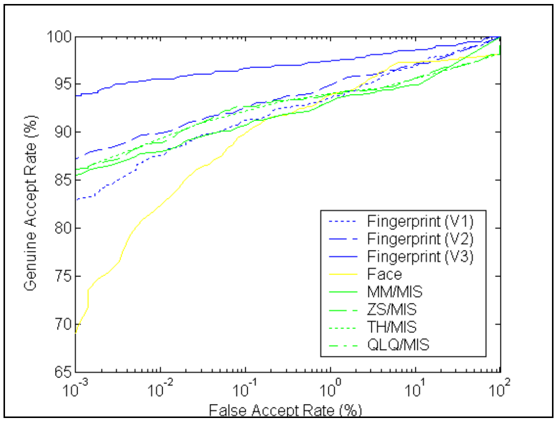


Fig. 6. ROC curves for MIS, normalization varied.

For MW fusion (Fig. 8), the matcher weights, calculated according to (1), are: 0.2, 0.22, 0.37 and 0.21, for the fingerprint matchers and the face matcher, respectively. From Figures 5-9 and Table 1, we see that QLQ and MM normalization methods lead to best performance, except for MIS fusion. Between these two normalization methods, QLQ is better than MM for fusion methods MAS and UW; and about the same as MM for the others.

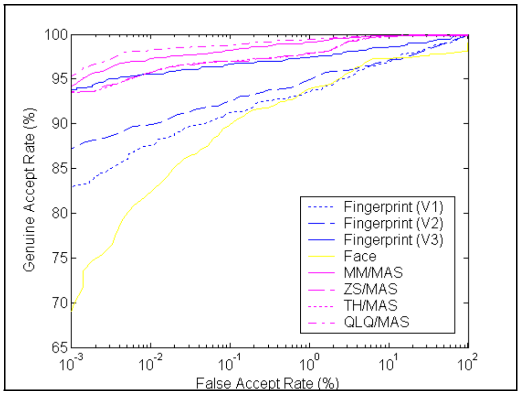


Fig. 7. ROC curves for MAS, normalization varied.

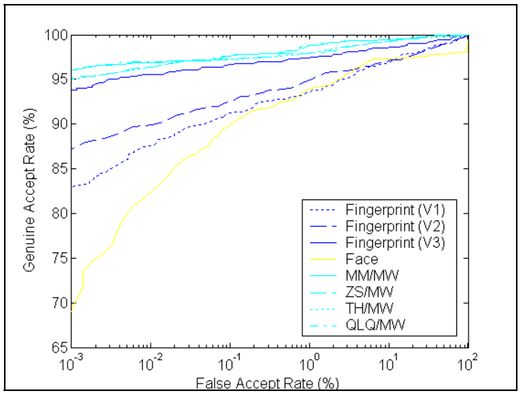


Fig. 8. ROC curves for MW, normalization varied.

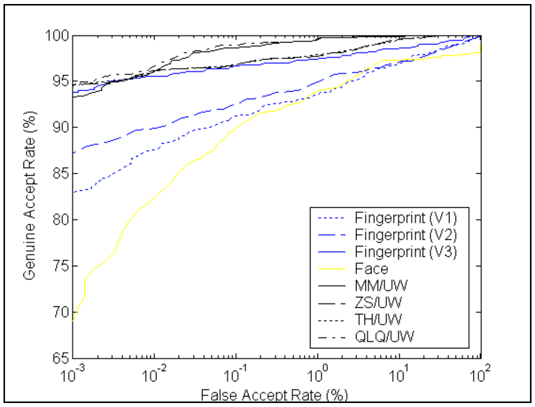


Fig. 9. ROC curves for UW, normalization varied.

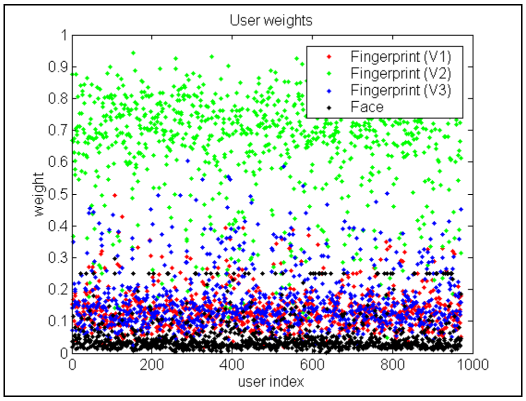


Fig. 10. Pictorial representation of user weights, for QLQ normalization.

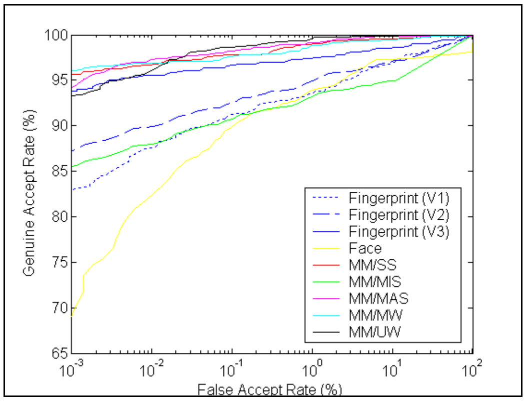


Fig. 11. ROC curves for MM, fusion varied.

4.4. Fusion

Figures 11-14 show the effect of each fusion method on system performance for different (but fixed) normalization methods. The ROCs for the individual fingerprint matchers and the face matcher are also shown for better comparison.

From Figures 11-14 and Table 1, we see that fusion methods SS, MAS and MW generally perform better than the other two (MIS and UW). But for the FAR range of [0.01%, 10%], UW fusion is better than the others. One reason that below 0.01% FAR the performance of UW fusion drops may be the estimation errors become dominant, since we have only one sample available for replacing the individual genuine distributions.

Note that parameter update (for normalization and/or fusion methods) can be employed for addressing the time varying characteristics of the target population. For example, the matcher weights can be updated every time a new set of EER figures are estimated; the user weights can be updated if the fusion system detects changes in the vulnerability of specific users, due to fluctuations in their *lambness*, etc.

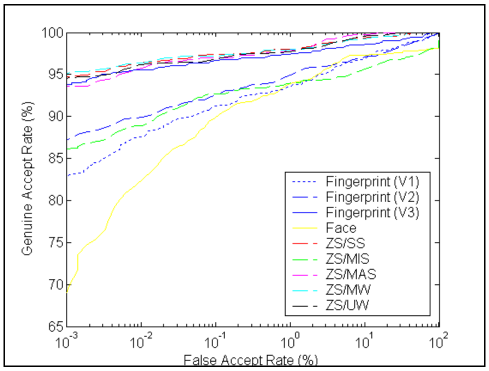


Fig. 12. ROC curves for ZS, fusion varied.

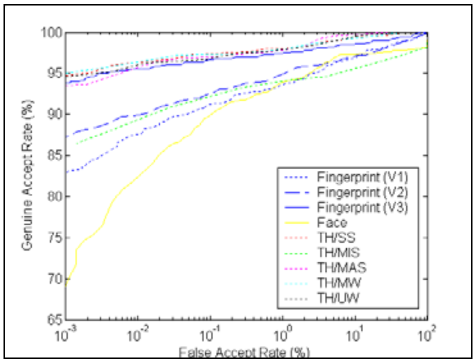


Fig. 13. ROC curves for TH, fusion varied.

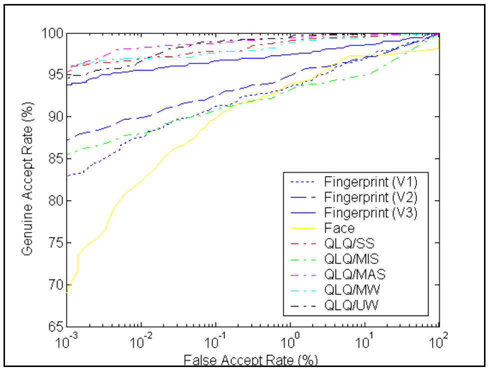


Fig. 14. ROC curves for QLQ, fusion varied.

4.5. Fusing Subsets of Matchers

ROC curves were generated for fusing subsets of the total matcher set. Here, we fixed the normalization method to QLQ and the fusion method to SS.

In Fig. 15 we see that fusing just the three fingerprint matchers (V1V2V3, with EER of 1.94%) is not as good as fusing all the available four matchers (V1, V2, V3 and Face) using QLQ/SS (see Figs. 5 and 14). This implies that even though the face matcher is not as good as any of the individual fingerprint matchers, it still provides complementary information for fusion.

Fusing individual fingerprint matchers separately with the face matcher (V1-Face, V2-Face, V3-Face; with EERs of 1.68%, 1.46% and 2.02%, respectively) we see that V2-Face performs better than V3-Face fusion. Since V3 is the better fingerprint matcher for our dataset, this result may seem counterintuitive. In fact this shows that the face matcher is best complemented with the V2 matcher, i.e., they make independent mistakes; whereas face matcher and V3 matcher make relatively more correlated mistakes.

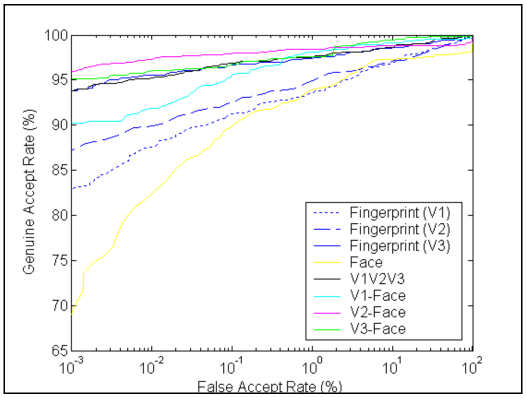


Fig. 15. Fusing subsets of matcher set.

4.6. Performance Variability

We are interested in determining how the performance of the fused system changes when using (i) an increasingly larger database, (ii) different equal-size databases, and (iii) many randomly assigned virtual subject databases.

Scalability. We created five new user databases from subsets of our original 972 user database: (i) the first 20% of all the users (194 users), (ii) the first 40% of all the users (389 users), (iii) the first 60% of all the users (583 users), (iv) the first 80% of all the users (778 users) and (v) 100% of all the users (972 users). Fig. 16 shows the associated ROC curves for an MM/SS based multimodal system using these datasets. The EERs corresponding to these five sets are 0.42%, 0.75%, 0.67%, 0.8%, and 0.99%, respectively.

We observe that the performance initially drops, but then quickly converges. For this relatively large, but limited, dataset we are unable to draw any general conclusions. It is widely believed that performance decreases as the database size increases. A possible explanation for this belief is that as the state space becomes more populated, differentiation within it, or some clustered areas, becomes more difficult. Another viewpoint is that performance trends cannot be extrapolated to larger populations. Thus a representative database of the intended size may be necessary to predict performance.

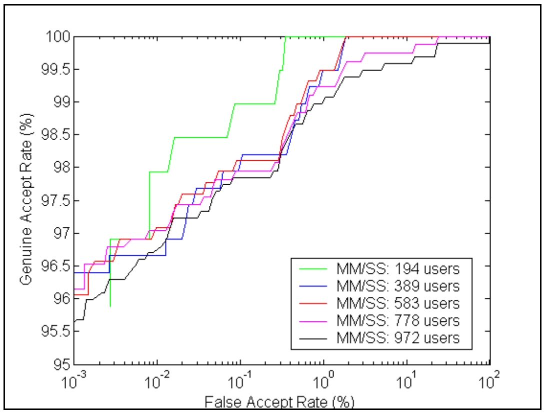


Fig. 16. Scalability: ROC curves for overlapping portions of the whole database.

Generalizability. We created two new user databases of 486 users each from *disjoint* subsets of the original database of 972 users. Fig. 17 shows the associated ROC curves for an MM/SS based multimodal system using these disjoint datasets. The EERs corresponding to these datasets are 0.68% and 1.45%, respectively. We see that the portion of the ROC curves above 0.4% FAR, show a considerable performance difference. Although we can draw no general trends, this implies that its necessary to use a representative database when determining expected performance, but there are presently no clear measurements/methods to determine if a database is representative. Similar results have been reported for performance variation of unimodal systems in [9].

Virtual Subjects. As explained previously, it is common practice to create virtual subjects in multimodal experiments. In our previous experiments, we consistently assigned a "physical finger" to a "physical face" to create a virtual subject. In this section, we randomly created 1000 virtual user sets (i.e., we randomly assigned the 972 face samples to the 972 fingerprint samples, 1000 times). In Fig. 18, we plot the ROC's for all of these 1000 cases, with the one used previously in this paper highlighted in red.

The minimum, mean, maximum and standard deviation of the EER set (with 1000 members) is found to be 0.82%, 1.1%, 1.5% and 0.11, respectively. The EER for the one case used previously in this paper is 0.99%. The close clustering of these curves, and the low standard deviation, supports the independence assumption between face and fingerprint biometrics and would seem to validate the use of virtual subjects. Furthermore the "thickness" of this cluster of curves supports other observations that performance estimates vary by nearly +/- 1%.

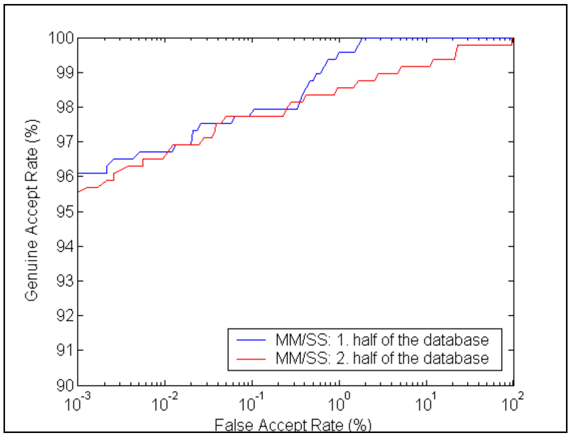


Fig. 17. Generalizability: ROC curves for disjoint portions of the whole database.

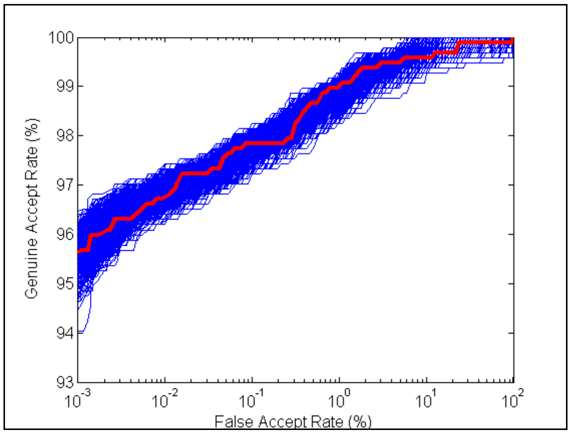


Fig. 18. Effects of virtual subject creation.

5. Conclusions

We examined the performance of multimodal biometric authentication systems using state-of-the-art Commercial Off-the-Shelf (COTS) fingerprint and face biometrics on a population approaching 1000 individuals, 10 times larger than previous studies. We introduced novel normalization and fusion methods along with well-known methods to accomplish match score level multimodal biometrics. Our work shows that COTS-based multimodal fingerprint and face biometric systems can achieve better performance than unimodal COTS systems. However, the performance gains are smaller than those reported by prior studies of non-COTS based multimodal systems (a ~2.3% gain here as compared to a ~12.9% gain reported in [2], at 0.1% FAR). This was expected, given that higher-accuracy COTS systems leave less room for improvement. Our analysis of fusion and normalization methods suggests that for authentication applications, which normally deal with open populations (e.g., airports) whose specific information is not known in advance, Min-Max normalization and Simple-Sum fusion generally out perform unimodal biometrics. For applications which deal with closed populations (e.g., a laboratory), where repeated samples and their statistics can be accumulated, our novel QLQ *adaptive normalization* and UW fusion methods tend to out perform Min-Max normalization and Simple-Sum fusion.

Our analysis of multimodal face-fingerprint pair systems shows that better performance is obtained by combining complementary systems rather than the best individual systems. And our investigations of performance variability across different datasets have provided evidence that the use of virtual subjects is valid, and offer initial estimates of variability for COTS-based multimodal systems .

6. References

[1] NIST report to the United States Congress, “Summary of NIST Standards for Biometric Accuracy, Tamper Resistance, and Interoperability”, November 13, 2000.

[2] R. Snelick, M. Indovina, J. Yen, A. Mink, "Multimodal Biometrics: Issues in Design and Testing", Proc. of The 5th International Conference on Multimodal Interfaces (ICMI 2003), November 2003, Vancouver, British Columbia, Canada.

[3] A.K. Jain, R. Bolle, and S. Pankanti, Eds. Biometrics: Personal Identification in Networked Society, Kluwer Academic Publishers, 1999.

[4] A. Ross and A.K. Jain, “Information Fusion in Biometrics”, Proc. of AVBPA, Halmstad, Sweden, June 2001, pp. 354-359.

[5] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, “Sheeps, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation”, Proc. of ICSLD 98, Sydney, Australia, November 1998.

[6] P.J. Huber, Robust Statistics, Wiley, 1981.

[7] A.K. Jain and A. Ross, “Learning User-Specific Parameters in Multibiometric System”, Proc. of International Conference on Image Processing (ICIP), Rochester, NY, September 2002, pp. 57-60.

[8] R.M. Bolle, S. Pankanti, and N.K. Ratha, “Evaluation techniques for biometrics-based authentication systems (FRR)”, Proc. of ICPR 2000, 15th International Conference on Pattern Recognition, Sept 2000, vol. 2, pp. 831 -837.

[9] P.J. Phillips, P. Grother, R.J. Micheals, D.M. Blackburn, E. Tabassi, and M. Bone, “Face Recognition Vendor Test 2002, Evaluation Report”, March 2003, ftp://sequoyah.nist.gov/pub/nist_internal_reports/ir_6965/FRVT_2002_Evaluation_Report.pdf

INTEGRATING PALMPRINT WITH FACE FOR USER AUTHENTICATION

Ajay Kumar, David Zhang
Department of Computing,
The Hong Kong Polytechnic University, Hong Kong.
Email: {*csajaykr, csdzhang*}*@comp.polyu.edu.hk*

ABSTRACT

This paper presents a new method of personal authentication using face and palmprint images. The facial and palmprint images can be acquired by using touchless sensors and integrated to achieve higher confidence in personal authentication. This has been confirmed from the qualitative and quantitative results shown in this paper. The proposed method of fusion uses a feed-forward neural network to integrate individual matching scores and generate a combined decision score. The significance of the proposed method is more than improving performance of bimodal system. Our method uses claimed identity of users as a feature for fusion. Thus the required weights and bias on individual biometric matching scores are automatically computed to achieve the best possible performance. The method proposed in this paper can be extended for any multimodal authentication system to achieve higher performance.

1. INTRODUCTION

The technology for trusted *e*-security is critical to many business and administrative process. There has been a newfound urgency after September 11 attacks to develop cutting-edge security technologies. However, the performance of currently available technology is yet to mature for its broad deployment in real environments. The biometrics-based characteristics *i.e.*, face, palmprint, iris, hand geometry, *etc*., are distinctive, cannot be forgotten or lost, and requires physical presence of the person to be authenticated [1]. Thus biometrics-based personal authentication systems are more reliable, convenient, and efficient than the traditional identification methods. The financial risks in personal authentication are high; the double dipping[*] in social welfare schemes are estimated around $40 billion and 40-80% of IT help desk calls are attributed to forgotten passwords [2]. Secure access control helps to minimize the security/terrorist threats at airports, airplanes, and security installations and is more relevant in the current world scenario.

The multimodal biometrics system allows integration of two or more biometric in order to cope up with the stringent performance requirements imposed for high security access. Such systems offer high reliability due to the presence of multiple piece of evidence and are vital for fraudulent technologies as it is more difficult to simultaneously forge multiple biometric characteristics than to forge a single biometric characteristic. One of the recent research problems in the design of multimodal biometrics system concerns with information fusion, *i.e*. how the individual modalities should be combined to minimize errors and achieve high accuracy.

1.1. Prior Work

Multimodal biometric systems have recently attracted the attention of researchers and some work has already reported in the literature [3]-[11]. Hong and Jain [3] combined fingerprint and face to achieve major improvement while Ben-Yacoub *et al*. [4] demonstrated this by integrating face with voice. Chatzis *et al*. [5] have used fuzzy clustering algorithm for the decision level fusion in personal authentication. Recently, bimodal biometric systems using face and iris [6], palmprint and hand-geometry [8], have shown promising results. Osladciw *et al*. [9] have presented a framework for multimodal biometric system that is adaptively tuneable to the security needs of user. Verlinde *et al*. [10] achieve decision level fusion by using parametric and non-parametric classifiers. Kittler *et al*. [11] have shown that the sum rule is most resilient for the estimation of errors in biometric fusion.

1.2. Proposed System

This paper investigates a bimodal biometric system using face and palmprint. Face has highest user acceptance and its acquisition is most convenient to users [12]. People have lot of concerns about hygiene, especially due to

[*] where an individual unlawfully benefits under multiple identity.

recent spread of SARS[†], while using biometric sensors *e.g*. fingerprint sensors. However the face and palmprint images can be conveniently acquired from the touchless sensors such as digital camera. One of the important features that is only available in personal authentication, but not in recognition, is the claimed user identity. The claimed user identity is unique for every user and can be used to restrict the decision space, *i.e*. range of matching scores, in user authentication. The claimed user identity can be suitably coded and then used as a feature to classify the genuine and impostor matching scores and is investigated in this paper The main contributions of this paper are twofold; (i) investigate a new bimodal biometric authentication system by integrating face and palmprint features, and (ii) propose a new decision level fusion strategy that uses claimed identity as a feature to classifier.

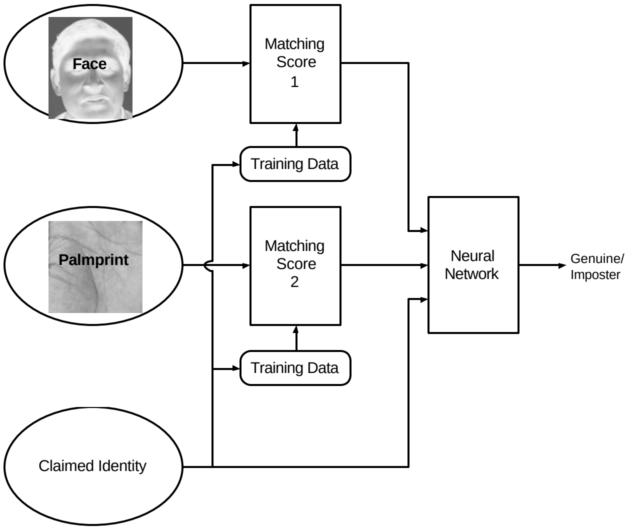


Figure 1: Personal Authentication using Face and Palmprint.

2. METHODOLOGY

The block diagram of the proposed bimodal biometric authentication method is shown in Figure 1. The acquired grey-level images from the palmprint and face are presented to the system. In addition, each of the users also presents its claimed identity. The matching scores from each of the two biometric are presented to a neural network classifier. As shown in Figure 1, the claimed user identity is also used as a feature to neural network classifier. The weights and bias of individual biometrics are automatically computed during the training of neural network. The trained neural network generates the combined decision scores for the claimed user identity and assigns them in one of two classes *i.e*. genuine or impostor.

3. FACE MATCHING

Several face recognition algorithms have been proposed in the literature [13]. Among these, the appearance based face algorithms are most popular and have been installed in real-environments [14]. The appearance based face authentication approach used in this work employed eigenfaces [15]. Each of the $M \times N$ grey-level face images from every subject are represented by a vector of $1 \times MN$ dimension using row ordering. The normalized set of such training vectors is subjected to principal component analysis (PCA). The PCA generates a set of orthonormal vectors, also known as eigenfaces, which can optimally represent the grey-level information in the training dataset. The projection of subjects training face image on eigenfaces is used to compute the characteristic features. The matching score for every test face image is generated by computing the similarity score between the feature vectors from the claimed identity (x_c) and computed characteristic feature vector (x_q).

$$h_{face} = \frac{\sum x_q x_c}{\sqrt{\sum x_q^2 \sum x_c^2}} \qquad (1)$$

4. PALMPRINT MATCHING

Palmprint contains several complex features, *e.g*. minutiae, principal lines, wrinkles and texture, which have been suggested for personal identification. The palmprint matching approach used in this work is same as detailed in [8]. Four directional spatial masks are used to capture line features from each of the palmprint images. The combined directional map is generated from voting of the resultant four images. The standard deviation of pixels, from each of the 24 × 24 pixel overlapping block with 25% overlap, in the combined image is used to form characteristic feature vector. The palmprint matching scores are generated by computing the similarity measure h_{palm}, similar to (1), between the feature vectors from acquired image and those stored during the training phase.

5. DECISION LEVEL FUSION USING NEURAL NETWORKS

Decision level fusion that can consolidate the decision scores from multiple evidences has shown [3]-[8], [16]-[17] to offer radical increase in performance. The genuine and impostor matching scores from the face and palmprint are used to train a feed-forward neural network (FFN). Our

[†] Severs Acute Respiratory Syndrome (SARS) is highly infectious disease prone to human touch.

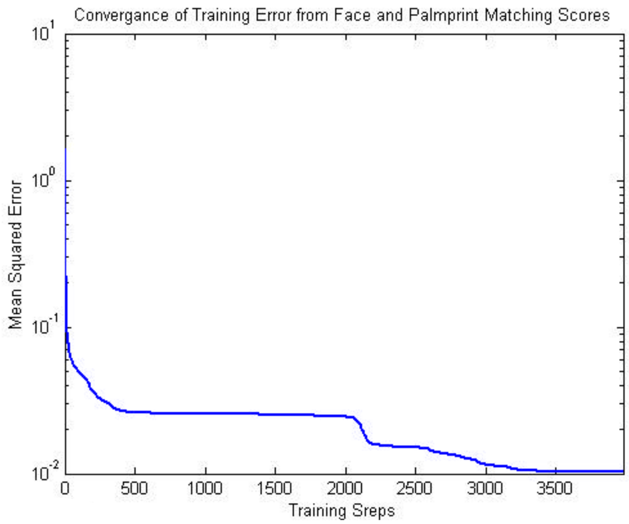


Figure 2: Convergence of training error from the Palmprint and Face matching scores.

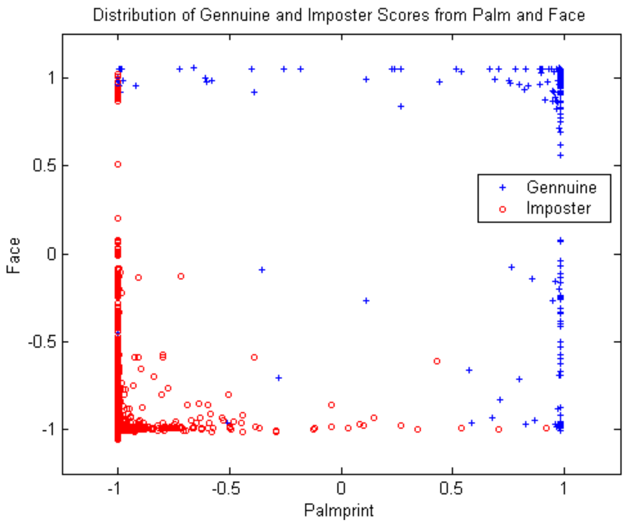


Figure 3: Distribution of genuine and imposter scores from the two biometric.

strategy is to use the claimed identity of every user as a feature to FFN. Execution speed of multi-layer feed-forward neural network is among the fastest of all models currently in use. Therefore this network may be the only practical choice for online personal authentication. A (FFN) with P_l neurons in the l^{th} (l = 1, …, Q) layer is based on the following architecture [18]:

$$j_j^l = \sum_{i=1}^{P_{l-1}} w_{ij}^{l-1,l}\, y_i^{l-1}, \quad y_j^l = g(j_j^l) \qquad (2)$$

where the sum of weighted inputs for j^{th} $(j = 1, \ldots, P_l)$ neuron in the l^{th} layer is represented by j_j^l. The weights from the i^{th} neuron at $(l-1)^{\text{th}}$ layer to the j^{th} neuron in the l^{th} layer are denoted by $w_{ij}^{l-1,l}$ and y_j^l is the output for j^{th} neuron in the l^{th} layer. The values -1 and 1, corresponding to 'impostor' and 'genuine' responses, were given to the three layers FFN during training as the correct output responses for expected classification during the training. The hyperbolic tangent sigmoid activation function was empirically selected for first two layers, while a linear activation function was chosen for third layer.

$$g(j_j^l) = tanh(j_j^l) \quad \text{for } l = 1, 2. \text{ and}$$
$$g(j_j^l) = a(j_j^l) \quad \text{for } l = 3. \qquad (3)$$

The back-propagation training algorithm is used for minimizing training function T_e defined by:

$$T_e = \frac{1}{\text{KP}_\text{Q}} \sum_{k=1}^{\text{K}} \sum_{j=1}^{\text{P}_\text{Q}} (y_{j,k}^{\text{Q}} - o_{j,k})^2 \qquad (4)$$

where k is an index for input-output pair and $(y_{j,k}^{\text{Q}} - o_{j,k})^2$ is the squared difference between the actual output value at the j^{th} output layer neuron for pair k and the target output value. The connection weights $w_{ij}^{l-1,l}$ are updated after presentation of every feature vector using a constant learning rate. The weights are updated using Levenberg-Marquardt algorithm [19] for faster convergence rate.

6. EXPERIMENTS AND RESULTS

The proposed method was investigated on available face database [20] from 40 subjects with 10 images per subject. The hand images from 40 subjects, with 10 images per subject, were acquired by using a digital camera. Each of the subjects for palmprint and face were randomly paired[‡] to obtain a bimodal set for every subject. The 300 × 300 region of interest, *i.e.* palmprint, were automatically segmented and features vectors of size 1 × 144 were extracted as detailed in [8]. Each of the 92 × 112 pixel face image was used to obtain 1 × 40 characteristic feature vector from the 40 eigenfaces. The matching scores for face and palmprint were computed by similarity measure (1). In this work, the first four images samples, from face and palmprint, were used for training and rest six were for testing. Thus genuine and impostor matching scores from the training samples were used to train 18/5/1 neural network as shown in Figure 1. The learning rate was fixed at 0.01 and the convergence of training error is shown in Figure 2. There is no guarantee that the achieved training error is global and therefore the FFN was trained 10 times with the same parameters and the result with the smallest

[‡] The mutual independence of biometric modalities [21] allows us to augment two biometric indicators that are collected individually.

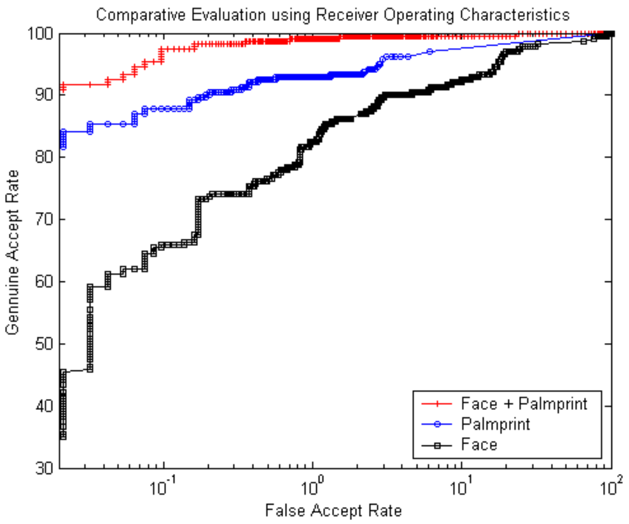


Figure 4: Comparative performance fore user authentication using Palmprint and Face.

of training errors of all the results are reported. The trained neural network was used to test 240 (40×6) genuine and 9360 (40×39×6) impostor matching scores from the test data.

The distribution of decision scores from trained neural network, from the test data, is shown in Figure 3. It can be seen that the two matching scores are quite distinct and separable by any two class linear discriminant function. The receiver operating characteristics for (i) face, (ii) palmprint, and (iii) using fusion of face and palmprint is shown in Figure 4. All these cases shown in Figure 4 employ the claimed identity as a feature to FFN. The variation of False Accept Rate (FAR) and False Reject Rate (FRR) with decision threshold for combined decision is shown in Figure 5. The cumulative distribution for combined impostor and genuine decision scores is shown in Figure 6.

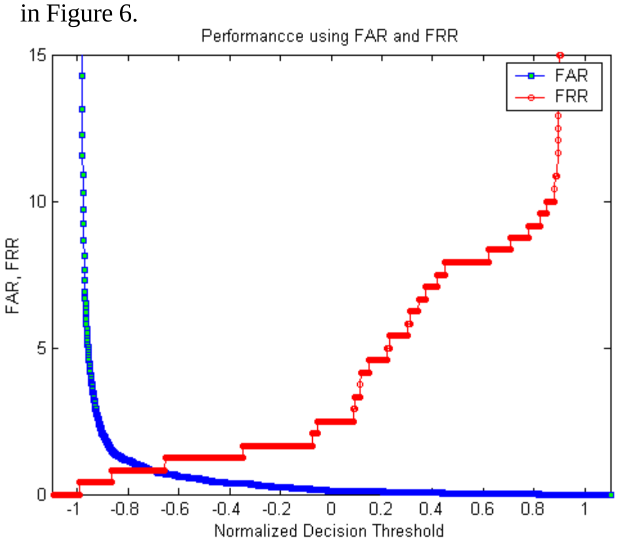


Figure 5: Variation of FAR and FRR scores with decision threshold for combined decision.

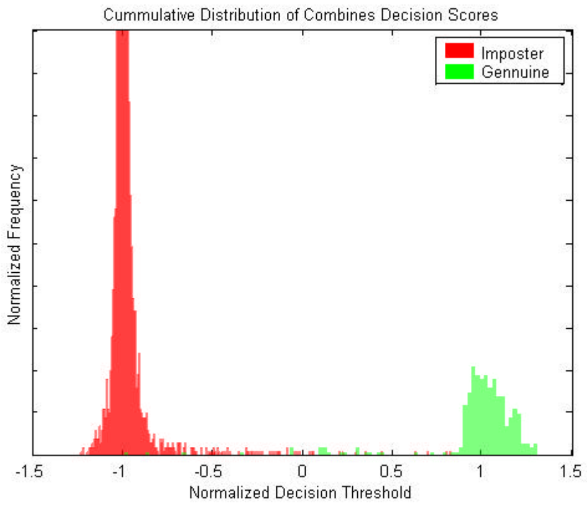


Figure 6: Comparative performance fore user authentication using Palmprint and Face.

The FAR and FRR scores for three distinct cases using total[§] minimum error (*TME*) is shown in Table 1. It is worth to mention that the total minimum error for the fusion was 2.80% when the claimed user identity was not utilized and 1.54% when claimed user identity was employed to train/test the FFN. In order to ascertain the improvement (or degradation) in the separation of genuine and impostor decision scores for the fusion, the performance index using three objective functions [22], were considered.

$$J_1 = \frac{\mu_g}{\mu_i}, J_2 = \frac{(\mu_g - \mu_i)^2}{\mu_g \mu_i}, J_3 = \frac{(\mu_g - \mu_i)^2}{s_g^2 + s_i^2} \qquad (5)$$

where m_g, m_i are the mean and s_g, s_i are the standard deviation of genuine and impostor distributions respectively. The scores for above performance indices were computed from the test data (Figure 6 scaled to positive axes) and are displayed in Table 2. The bracketed entries in this table show the respective scores when the claimed identity of user is not utilized. These entries can be used to interpret the performance increase when the claimed user identity is used as a feature. Table 2 also shows the equal error rate (*EER*) for each of the corresponding cases.

Table 1: Performance scores for total minimum error.

	FAR	FRR	Decision Threshold
Face	3.04	10	-0.71
Palmprint	3.75	3.15	-0.99
Fusion at Decision	0.70	0.83	-0.66

[§] Sum of FAR and FRR for the combined decision.

Table 2: Performance indices from the experiments.

	J_1	J_2	J_3	*EER*
Face	3.85 (1.05)	2.11 (0.002)	4.42 (2.34)	8.33 % (8.69 %)
Palmprint	4.38 (1.03)	2.61 (0.001)	8.61 (3.71)	3.65 % (4.32 %)
Fusion at Decision	**4.84** **(4.78)**	**3.04** **(2.988)**	**35.57** **(23.78)**	**0.84** % **(2.09 %)**

7. CONCLUSIONS

The grey-level images of palmprint and face can be simultaneously acquired and used to achieve the performance that may not be possible by single biometric alone. The performance improvement shown in Figure 4 confirms the usefulness of the proposed bimodal system. Furthermore, this can also be quantitatively ascertained from the results shown in Table 2. All the three performance indicators, *i.e.* J_1, J_2, and J_3 have shown improvement when both the biometrics are utilized. However, the scores from index J_3 are substantially higher than those from J_1 or J_2. This is due to the fact that J_3 also accounts for the standard deviation of decision scores. Therefore J_3 can be used as a reliable measure to evaluate the performance in biometrics. The performance scores in first two rows of Table 2 also suggest that the claimed user identity has significant effect in improving performance even for unimodal authentication, *i.e.* face and palmprint. This improvement is attributed to the fact that the FFN classifier uses the claimed user identity to reduce the decision space, *i.e.* range of valid matching scores, for the corresponding user. The inclusion of claimed user identity can be used improve the performance in unimodal authentication systems without any extra cost and is therefore recommended.

The significance of the proposed method is more than improving performance for a bimodal system. Our method has utilized the claimed identity of subjects as a feature for fusion. The employed neural network thus automatically computes the weights and bias for the individual biometric matching scores to achieve the best possible performance. The performance scores shown in Table 2 suggest that this is indeed the case. In order to achieve more reliable estimate on the performance it is desirable to evaluate this method on significantly large database and we are currently working on this. A qualitative summary of related work on multimodal user authentication is presented in Table 3. The proposed method of fusion can be extended to any multimodal system to achieve higher performance. Additionally, the

Table 3: A summary of related work on multimodal user authentication.

Authors	**Biometric Modalities**	**Fusion Strategy**	**Performance Criteria**	**Touchless Sensors**
Hong and Jain [3]	Face, Fingerprint	Hierarchal decision using combined imposter distribution	FRR, ROC	No
Ben-Yacoub *et al*. [4]	Voice, Face	SVM, Bayes	FAR, FRR	Yes
Chatzis *et al*. [5]	Voice, Face	FVQ, RBF	FAR, FRR	Yes
Wang *et al*. [6]	Face, Iris	User-specific RBF, Weighted Sum Rule	*TME*	Yes
Kumar *et al*. [8]	Palmprint, Hand Geometry	Max Rule	*TME*, ROC	Yes
Jain and Ross [7]	Face, Fingerprint, Hand Geometry	User-specific threshold, Weighted Sum Rule	ROC	No
Kittler *et al*. [11]	Face, Face Profile, Voice	Sum, Max, Median, Product Rule	*EER*	Yes
Kumar and Zhang	Face, Palmprint	FFN based fusion with user claimed identity	J_1, J_2, J_3	Yes

performance improvement in multimodal system can be also be ascertained by two class separation functions (5), rather than just ROC or total minimum error as used in prior work [3]-[8].

8. REFERENCES

[1] D. Zhang (*ed*.), *Biometrics Solutions for Authentication in an e-World*, Kluwer Academic Publishers, USA, 2002.

[2] Gartner, Inc., http://www.gartner.com

[3] L. Hong and A. K. Jain, “Integrating faces and fingerprints for personal identification,” *IEEE Trans. Patt. Anal. Machine Intell*., vol. 20, pp. 1295-1307, Dec. 1998.

[4] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Trans. Neural Networks*, vol. 10, pp. 1065-1074, Sep. 1999.

[5] V. Chatzis, A. G. Bors, and I. Pitas, “Multimodal decision-level fusion for person authentication,” *IEEE Trans. Sys. Man Cybernetics*: *Part A*, vol. 29, pp. 674-680, Nov. 1999.

[6] Y. Wang, T. Tan, and A. K. Jain, “Combining face and iris for identity verification,” *Proc. AVBPA*, Guildford (U.K.), pp. 805-813, Jun. 2003.

[7] A. K. Jain and A. Ross, "Learning User-specific Parameters in a Multibiometric System,” *Proc. ICIP 2002*, Rochester, pp. 57-70, New York, Sep. 2002.

[8] A. Kumar, D. C. M. Wong, H. Shen, and A. K. Jain, "Personal verification using palmprint and hand geometry biometric,” *Proc. AVBPA*, pp. 668-675, Guildford, UK, June 2003.

[9] L. Osadciw, P. Varshney, and K. Veeramachaneni, “Improving personal identification accuracy using multisensor fusion for building access control applications,” *Proc. ISIF* 2002, pp. 1176-1183, 2002.

[10] P. Verlinde, G. Chollet, and M. Acheroy, “Multi-modal identity verification using expert fusion,” *Information Fusion*, vol. 1, pp. 17-33, 2000.

[11] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers," *IEEE Trans. Patt. Anal. Machine Intell*., vol. 20, pp. 226-239, Mar. 1998.

[12] S. Prabhakar, S. Pankanti, and A. K. Jain, “Biometric Recognition: Security and Privacy Concerns,” *IEEE Security & Privacy Magazine*, pp. 33-42, Mar./Apr. 2003.

[13] M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting faces in images: a survey,” *IEEE Trans. Patt. Anal. Machine Intell*., vol. 24, pp. 34-58, Jan. 2002.

[14] Visage Technology Inc., http://www.viisage.com

[15] M. A. Turk and A. P. Pentland, “Eigenfaces for Recognition,” *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71-76, 1991.

[16] S. Prabhakar and A. K. Jain, "Decision level fusion in fingerprint verification," *Pattern Recognition*., vol. 35, pp. 861-874, 2002.

[17] R. Brunelli and D. Falavigna, “Personal identification using multiple cues,” *IEEE Trans. Patt. Anal. Machine Intell*., vol. 17, pp. 955-966, Oct. 1995.

[18] A. Kumar, “Neural network based detection of local textile defects,” *Pattern Recognition*, vol. 36, pp. 1645-1659, 2003.

[19] Timothy Masters, *Advanced algorithms for neural networks*: *A C++ sourcebook*, Wiley, New York, 1995,

[20] The Olivetti Research Database of Faces; http://www.cam-orl.co.uk/facedatabase.html

[21] A. Ross and A. K. Jain, “Information fusion in biometrics,” *Pattern Recognition Lett*., vol. 24, no. 13, pp. 2115-2125, Sep. 2003.

[22] A. Kumar and G. Pang, “Defect detection in textured materials using optimized filters,” *IEEE Trans. Systems, Man, and Cybernetics*: *Part B, Cybernetics*, vol. 32, pp. 553-570, Oct. 2002

MULTI-MODAL FACE AND SPEAKER IDENTIFICATION ON A HANDHELD DEVICE

Timothy J. Hazen, Eugene Weinstein,
Ryan Kabir, Alex Park

MIT Computer Science and
Artificial Intelligence Laboratory
Cambridge, MA 02139, USA

Bernd Heisele

Honda Research Institute USA, Inc.
Boston, MA 02111, USA

ABSTRACT

In general, most systems for face and speaker identification are tested on high quality data collected in well-lit and quiet environments. In this study, we investigate the application of existing face and speaker identification techniques to the task of user authentication on a handheld device. In this context, the audio/visual capture hardware is of lower quality than equipment typically used in laboratory experiments. Additionally, variable background conditions which can degrade the audio/visual signal may be present. These factors can be expected to harm the performance of the system. Under these circumstances, using a combination of biometric modalities can improve the robustness and accuracy of the person identification task. In this paper, we present our approach for combining both face and speaker identification technologies on a handheld device, and experimentally demonstrate a fused multi-modal system which achieves a 90% reduction in equal error rate over the better of the two independent systems.

1. INTRODUCTION

This paper investigates the integration of two biometric techniques, face and speaker identification, into handheld devices. This research is spurred by the recent increased popularity of commercially-available handheld computers which have allowed computation to become more mobile and pervasive. Formerly specialized devices, such as cellular telephones, now offer a range of capabilities beyond simple voice transmission, such as the ability to take, transmit and display digital images. As these devices become more ubiquitous and their range of applications increases, the need for security also increases. To prevent impostor users from gaining access to sensitive information, stored either locally on a device or on the device's network, security measures must be incorporated into these devices. Face and speaker verification are two techniques that can be used in place of, or in conjunction with, pre-existing security measures such as personal identification numbers or passwords.

Handheld devices offer two distinct challenges for standard face and voice identification approaches. First, their mobility ensures that the environmental conditions the devices will experience will be highly variable. Specifically, the audio captured by these devices could contain highly variable background noises producing potentially low signal-to-noise ratios. Similarly, the images captured by the devices can contain highly variable lighting and background conditions. Second, the quality of the video and audio capture devices is also a factor. Typical consumer products are constrained to use audio/visual components that are both small and inexpensive, resulting in a lower quality audio and video than is typically used in laboratory experiments.

To examine these issues we have developed a prototype system for incorporating two biometric techniques, speaker identification and face identification, into a mobile device. Results of an early evaluation of this system were previously reported in [1]. In our previous study, we evaluated a combined face and speaker identification system within a user verification "login" scenario on an iPAQ handheld computer. The combined system was able to achieve a 50% reduction in the verification equal error rate (EER) over a system using only our speaker identification technology. This large improvement in performance was attained despite the fact that speaker identification system achieved an EER that was 75% smaller than that of the face identification system. This result was surprising because it showed that large improvements could be obtained through the combination of different biometric systems, even when one of the systems was vastly superior in accuracy to the other. In the work conducted in this paper, we improve upon our previous results by replacing our older, simpler face identification system with a newer state-of-the-art system.

The rest of the paper is organized as follows. We first present an overview of our two biometric techniques and the fusion technique for combining them. Next, we discuss the mobile-device paradigm in which we are conducting our experiments and the methods of data collection employed. We follow this with experimental results showing the performance of the two biometric techniques on the data we have collected, both individually and in combination. Finally, we summarize and discuss the results and present plans for future directions of our work.

2. PERSON IDENTIFICATION

2.1. Speaker Identification

Speech has long been recognized as a reasonable biometric for person identification. However, speech is a variable signal whose main purpose is not to specify a person's identity but rather to encode a linguistic message. In systems where the linguistic content of the speech is unknown (e.g. for surveillance tasks), text-independent speaker identification systems are generally used. However, in security applications where the user is cooperative in the attempt to prove his/her identity, the linguistic content of the speech message is typically known and can be tightly constrained. In this case, a text-dependent system can be used. When the linguistic content of the message is known, text-dependent speaker recognition systems generally perform better than text-independent systems because they can tightly model the characteristics of the specific phonetic-content contained in the speech signal.

A common technique used in speech-based person identification is to prompt the user with a randomly generated challenge phrase. During authentication, automatic speech recognition can be used to verify that the spoken utterance matches the prompted utterance. For this type of scenario, it is both reasonable and beneficial to use the automatic speech recognition (ASR) output to leverage the phonetic constraints that give text-dependent systems their advantage. In [2], two techniques were described that use the ASR output during the analysis of the phonetic content from the test utterance.

In our speaker adaptive ASR approach, the system uses speaker-dependent speech recognizers to model each speaker. During training, phonetically transcribed enrollment utterances are used to train context-dependent phonetic models for each speaker. During testing, a speaker-independent ASR component generates a phonetic transcription from the test utterance. This transcription is then used by the system to score each segment of speech against each speaker-dependent phonetic model. Modeling speakers at the phonetic level can be problematic because enrollment data sets are typically too small to build robust speaker-dependent models for every context-dependent phonetic model. To compensate for this difficulty, we use an adaptive scoring approach in which the speaker-dependent (SD) score is interpolated with a speaker-independent (SI) score.

Mathematically, if the word recognition hypothesis assigns each feature vector x from the utterance X to phonetic unit j, then the score for speaker S_i, $p(X|S_i)$, is given by

$$\frac{1}{|X|}\sum_{x\in X}\log\left(\frac{\lambda_{i,j}p_{SD}(x|M_j,S_i)+(1-\lambda_{i,j})p_{SI}(x|M_j)}{p_{SI}(x|M_j)}\right)$$

where M_j is the model for phonetic unit j and $\lambda_{i,j}$ is an interpolation factor given by

$$\lambda_{i,j}=\frac{n_{i,j}}{n_{i,j}+\tau}\ .$$

In this equation, $n_{i,j}$ is the number of training examples of phonetic unit j observed for speaker S_i, and τ is a global tuning parameter that is set empirically using a separate development set. The log ratio in the equation generates positive scores when the input speech is a good match to a particular speaker's models and negative scores when the speech is a poor match.

This scoring strategy results in models that capture detailed phonetic-level characteristics for a speaker when sufficient training data is available, but relies more on speaker independent models for phonetic units with sparse training data. Thus, for cases with limited training data, the speaker independent model provides a more *neutral* score. In the limiting case, if no speakers have training data for any of the phones observed in a particular test utterance, then they will all receive the same neutral score of zero, which is an intuitively consistent result.

2.2. Face Identification

The face identification framework used in our work is similar to the one described in [3], but with some differences in detection and classification methods.

2.2.1. Face Detection

A two-step process is used for face detection. First, a fast hierarchical classifier similar to the one described in [4] is applied to the captured image, to roughly localize the face in the image. The region around the face is then cropped out from the larger image, histogram equalized, and scaled to a fixed size.

Next, a component-based face detector [3] is applied to the extracted region to precisely localize the face and to detect facial components. This method first independently applies component detection classifiers to the face image. Each of these support vector machine (SVM) classifiers is trained to detect a particular component, such as a nose, mouth, or left eyebrow. In all, 14 face components are used, and each component classifier is evaluated over a range of positions in the vicinity of the expected location of the desired component. A geometrical configuration classifier is

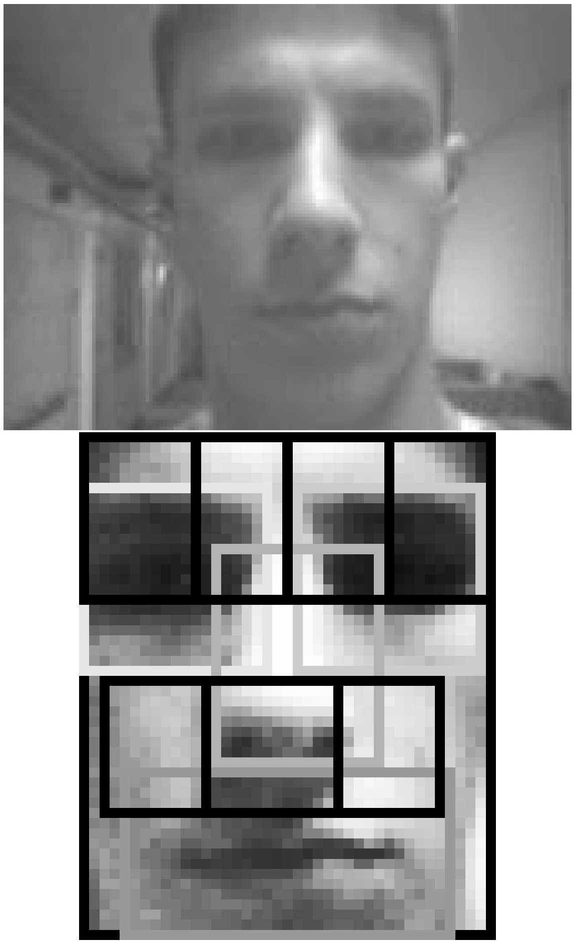


Fig. 1. A sample image and its face detection result with the face component regions superimposed.

then applied to the combined output of each of the 14 component classifiers from each candidate position. The candidate positions that yields the highest output of the second-level classifier are taken to be the detected component positions.

Ten out of the 14 components are used for face recognition. The remaining four are not used because they either overlap heavily with other components, or display few structures of use in distinguishing people from one another. The gray values of the ten selected components are normalized in size and combined into a single feature vector. The feature vector serves as input to the face recognizer. Figure 1 illustrates an enrollment image, as well as its selected face region with the positions of the facial components as detected by our system.

2.2.2. Face Recognition

For recognition, a one-vs-all SVM scheme is used, where one classifier is trained to distinguish each person in the database from all the others [5]. In the SVM training process, for each person's classifier, the feature vectors corresponding to that person's training images are used as positive examples, and the feature vectors corresponding to all the others' images are used as negative examples. The SVM training process finds the optimal hyperplane in the feature space that separates the positive and negative data points. Since the training data may not be separable, a mapping function corresponding to a second-order polynomial SVM kernel function [5] is applied to the data before training.

The runtime recognition process consists of computing the SVM classifier output score for each person's SVM classifier [5]. The scores are zero-centered – that is, a score of zero means the data point lies directly on the decision hyperplane, and positive and negative scores mean the data point lies on the positive and negative example side of the decision hyperplane, respectively. The absolute value of the SVM output is a multiple of the distance from the decision hyperplane, and could be normalized to produce the distance. Thus, a highly positive score represents a large degree of certainty that the data point belongs to the person the SVM was trained for, and a highly negative score represents the opposite. The output scores from all SVM classifiers make up the n-best list that we treat as our face recognition result.

For our face identification task, we collected and tested frontal face image data only. Most state of the art face identification systems attempt to account for rotations in and out of the image plane, and/or occlusions – which would be present in a typical surveillance task. However, for the handheld face identification problem, the user will be cooperating with the identification process; and in general, the user certainly will be looking at the screen of the handheld device as he or she is using it. Thus, accounting for heavily rotated or occluded faces is not important in this project. Generally, rotations or occlusions in face images make the problem of identification more challenging; thus, our problem is easier in this respect. Nonetheless, the variable lighting and background conditions and inexpensive camera present an orthogonal challenge, to ensure the nontriviality of our problem.

2.3. Multi-Modal Fusion

Past work on fusing face and speaker classifiers has generally used very simple combination strategies. Poh and Korczak used a logical AND rule on the results of their independent face and speaker systems [6]. This rule is most useful when the goal is to limit false acceptances, since both classifiers must accept the user in order to produce an acceptance by the fused-classifier. Brunelli and Falavigna [7] and Kittler *et al* [8] use basic probabilistic combination operators on the outputs from their independent recognizers. Bigün *et al* utilize a Bayesian statistics approach which compensates for biases and interdependencies between different classifiers [9]. An alternative to these statistical fusion approaches is the use of discriminatively trained methods such as decision trees or linear discriminant functions [10].

In our work, a linear weighted summation is employed for the classifier fusion where the weights for each classifier are trained discriminatively on a held-out development set using minimum classification error (MCE) training. The MCE training optimizes the equal error rate of false acceptances and false rejections under the user verification scenario. Because the final decision only requires the combination of two independent classifiers, only one additional parameter (the ratio of the weights of the classifiers) needs to be learned. A simple brute force sampling of the parameter space is used for this MCE training. More complicated techniques (such as gradient descent training) could be used in situations where more than two scores must be fused.

3. EXPERIMENTS

3.1. The Handheld Device

For our experiments we utilized a collection of iPAQ handheld computers. Speech data were collected utilizing the built-in microphone of the iPAQ. Two different models of iPAQs were used, with two different models of off-the-shelf, inexpensive electret condenser microphones. Face data were collected using a 640x480 CCD camera located on a custom-built expansion sleeve for the iPAQ. The iPAQ handheld computer, combined with the custom sleeve, is the handheld device platform used for pervasive computing research in the MIT Oxygen Project [11]. An image of the iPAQ with the expansion sleeve is shown in Figure 2. Because of the current computation and memory limitations of the iPAQ handhelds, the images and audio are captured by the handheld device, but then transmitted over a wireless network to servers which perform the operations of face detection, face identification, speech recognition, and speaker identification. In future work we hope to improve the computational efficiency and memory footprints of our systems so they can be deployed directly on small handheld devices.

3.2. The Login Scenario

Our experiments were conducted using a login scenario that combined face and speaker identification techniques to perform the multi-biometric user verification process. When "logging on" to the handheld device, users snapped a frontal view of their face, spoke their name, and then recited a prompted lock combination phrase consisting of three randomly selected two digit numbers (e.g. "*25-86-42*"). The system recognized the spoken name to obtain the "claimed identity". It then performed face verification on the face image and speaker verification on the prompted lock combination phrase. Users were "accepted" or "rejected" based on the combined scores of the two biometric techniques.

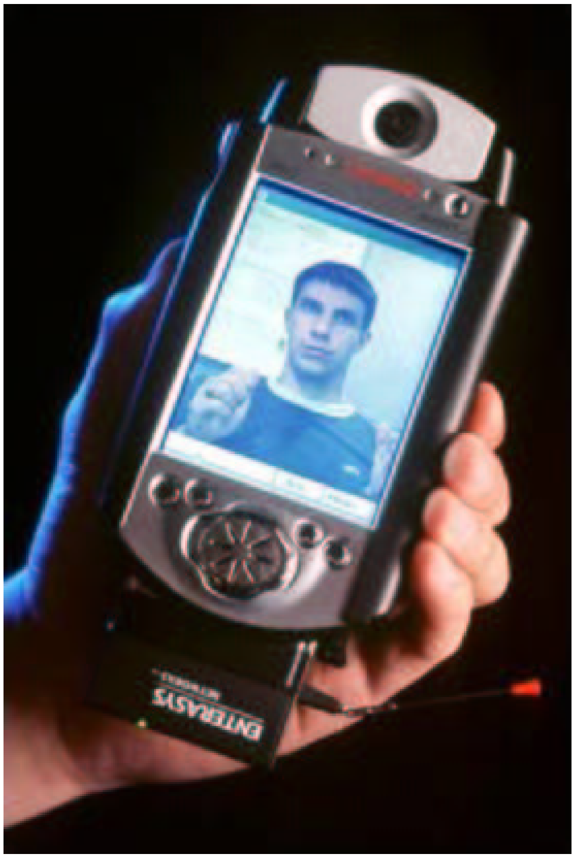


Fig. 2. The iPAQ handheld computer used in this study.

3.3. Data Collection

For our set of "enrolled" users, we collected face and voice data from 35 different people. Each person performed eight short enrollment sessions, four to collect image data and four to collect voice data. For each voice session, each user recited 16 prompted lock-combination phrases. Each image collection session consisted of the user taking 25 frontal face images in a variety of rooms in our lab with different lighting conditions. No specific constraints were placed on the distribution of the locations and lighting conditions; users were allowed to self-select the locales and lighting conditions of their images. To illustrate the quality of the images, Figure 3 shows two sample images captured during

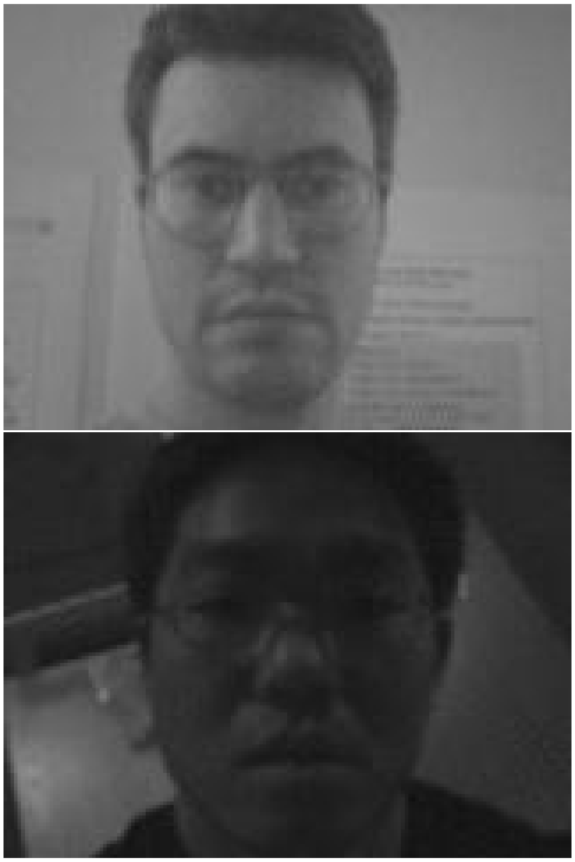


Fig. 3. Two sample images collected on the iPAQ.

the data collection.

During image collection, a fast face detector [12] was applied to each captured image to verify that the image indeed contained a valid face. This face detector occasionally rejected images when it failed to locate the face in the image with sufficiently high confidence. When this occurred the user was instructed to capture a new image. Due to a conservative face detection confidence threshold, no false positives (i.e., images with incorrectly detected faces) were observed from this face detector during the data collection. It is important to note that the face detector used during our data collection was not the same one used in the experiments in this paper.

Each voice and image session was typically collected on a different day, with the time span between sessions often spanning several days and occasionally a week or more. Each enrollment session typically lasted less than 5 minutes with the total enrollment time taking approximately 30 minutes on average. In total this yielded 100 images and 64 speech samples per enrolled user for training. An additional set of four enrollment sessions of audio data (i.e., 64 additional utterances) from 17 of the training speakers was available for development evaluations and multi-modal weight fusion training.

For our evaluation, we collected 16 sample login sessions from 25 of the 35 enrolled users. This yielded 400 unique utterance/face evaluation pairs from enrolled users. We also collected 10 impostor login sessions from 20 people not in the set of enrolled users for an additional 200 utterance/face evaluation pairs from unenrolled people.

We used the evaluation data to perform our user verification experiments. Each utterance/face pair from in-set speakers was used as a positive example of that user. This yielded a total of 400 positive examples for our evaluation. Each utterance/face pair from each in-set user could also be used as an impostor for the other 34 users in the enrolled set. This yielded 13600 impostor examples from in-set speakers. Each utterance/face pair collected from out-of-set impostors was also used to generate an impostor example for each of the 35 users in the enrolled set. This yielded 7000 impostor examples from users not in the enrollment set. In general, it is expected that impostors that have never been observed by the system will generate more classification errors than enrolled users who try to impersonate other enrolled users. This is because the models are trained to discriminate between users observed in the training data and thus may not generalize well to unseen users.

3.4. Training

The face and speaker systems were trained on the enrollment data for the 35 enrolled users. To train the fusion weights, one of the four face enrollment sessions was held out and a development face ID system was trained on the remaining three face sessions. Face identification scores from this held-out set were pairwise combined with speaker identification scores generated for utterances from the existing speaker identification development set. The true in-set examples and in-set impostor examples were provided to the MCE weight training algorithm previously described to generate the multi-modal fusion weights.

3.5. Face Detection Issues

For the experiments presented in this paper, the face detection algorithm used during the evaluation is not the same as the face detection algorithm used during the data collection. The detection algorithm used during the evaluation was specifically tuned to accept facial images that are well suited to the component-based classification method used for face identification. Because this classification method works best with frontal images of faces that are not tilted or contorted, the face detection algorithm was initially tuned such that tilted or contorted faces were rejected. The face detection algorithm used during our data collection was less conservative in its accept/reject decision of a hypothesized

Table 1. User verification results expressed as equal error rates (%), when forcing the face detector to output a detected face, on three systems (face only, speaker only, and multi-modal fusion) under two impostor conditions (known in-set impostors vs. unknown out-of-set impostors).

System	In-set Impostors	Out-of-set Impostors
Face	3.21%	4.87%
Speaker	0.75%	1.66%
Fused	0.24%	0.66%

face in an image. As a result, a sizable number of images in the training and evaluation data sets were rejected by the new face detection algorithm.

Because of the reduced number of images for our evaluation, we could not make a direct comparison with our previous test results. To allow us to make this comparison, we elected to run two experiments, one where the conservative face-detection decisions were used and a second experiment where the face detection algorithm was forced to output a detected face even if the image's detection score fell below the standard acceptance threshold. These two experiments allow us to examine the trade-off between the added gain in accuracy enabled by stricter control in the input facial images, and the potential added inconvenience of requiring users to provide an untilted, uncontorted frontal image.

3.6. Experimental Results

3.6.1. Forced Face Detection Results

Table 1 shows our user verification results for three systems (face ID only, speaker ID only, and our full multi-modal system) under two different impostor conditions (using only known in-set impostors vs. using only unknown out-of-set impostors). This experiment uses a detection threshold which forces the face detector to output a face hypothesis for all of the images, even when the detection confidence score is low. Figure 4 shows the results for the out-of-set impostor evaluation on a detection error trade-off (DET) curve.

Several observations should be made from these results. First, the speaker ID system has an equal error rate (EER) which is three times smaller than that of the face ID system when evaluated with unknown out-of-set impostors. These face ID results are better than our previously reported results in which the face ID system produced an EER which was four times larger than the speaker ID EER.

Next, the combined system has a 60% reduction in EER from 1.66% in the speech only system to 0.66% in the combined system. This is a slightly better improvement than the 50% reduction we had observed in our previous study. This demonstrates that sizable improvements can be obtained when multiple independent biometric techniques are combined even when one biometric technique performs substantially better than others.

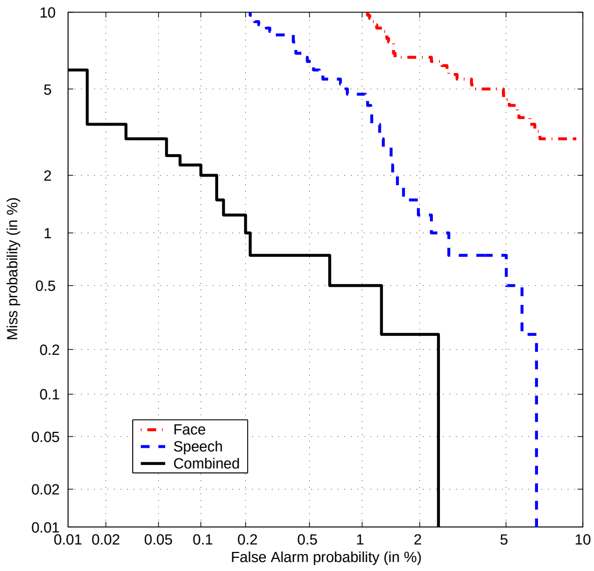


Fig. 4. DET curves for face and speech systems run independently and in combination when tested using impostors unknown to the system and when using a face detector that is forced to output a detected face for each input image.

Finally, it is interesting to note that the combined system achieves an EER of only 0.24% on the in-set impostor experiment. In other words, the EER when using the unknown impostors is 2.75 times greater than the EER of the in-set impostor experiment. This shows the importance of evaluating the system using people that are not part of the training data.

3.6.2. Conservative Face Detection Results

When applying the conservative face detection threshold to the evaluation utterances, 12% of the images were rejected. To evaluate the system under these conditions, the face ID system was first re-trained using the same threshold

Table 2. User verification results expressed as equal error rates (%), when using the conservative face detection threshold on three systems (face only, speaker only, and multi-modal fusion) under two impostor conditions (known in-set impostors vs. unknown out-of-set impostors).

System	In-set Impostors	Out-of-set Impostors
Face	1.66%	2.57%
Speaker	0.77%	1.63%
Fused	0.00%	0.15%

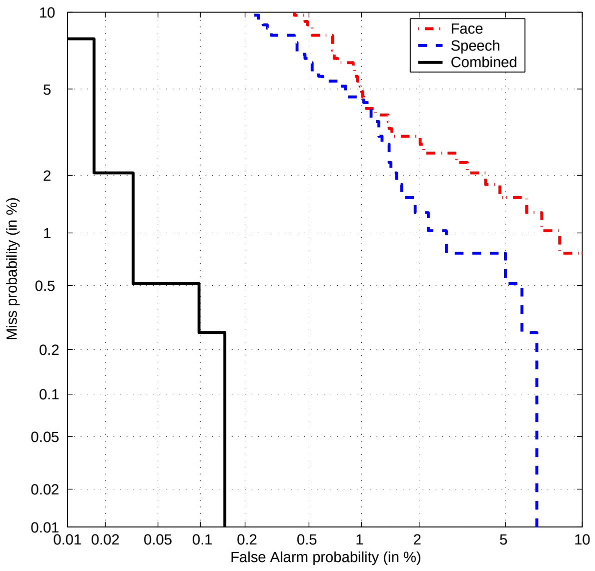


Fig. 5. DET curves for face and speech systems run independently and in combination when tested using impostors unknown to the system and when using the conservative face detection threshold.

for detection. The system's verification results were then re-computed using the 88% of the data that passed the more conservative face detection threshold.

Table 2 shows the equal error rates under these new constraints. The face ID system shows a nearly 50% improvement in EER performance over the forced detection result when the images with poor face detection scores were discarded. When used in conjunction with the speaker ID component, the combined system achieved an EER of only 0.15% when testing with out-of-set impostors. This a sizeable 90% reduction in EER from the speech only system! This combined system also achieved perfect separation between true users and in-set impostors resulting in a 0.0% EER on the in-set impostor experiment. This demonstrates that highly accurate biometric authentification can be obtained if the user is willing accept additional constraints on the verification process that may increase the inconvenience of the system. Unfortunately, because so few errors are observed, due to the limited size of our evaluation set, it is not possible to make any firm claims about the absolute level of the error rate of the system. We plan to increase the size of our evaluation set in future experiments.

3.6.3. Comparison with YOHO Corpus

To examine the degradation that might be experienced when our speaker identification technique is utilized in a mobile environment, we compared the performance of closed-set speaker recognition on the mobile handheld data set against the performance of our system on the tightly constrained YOHO corpus, which uses the same lock combination phrase approach that we employed [13]. It is important to note that the YOHO corpus was collected using a single close-talking telephone handset in a quiet office, and thus does not suffer from the degradations that are present in our mobile devices due to the low quality far-field microphone and the variable background conditions. In [2], it was shown that our system's speaker recognition error rate was 0.31% over YOHO's closed-set of 138 speakers. Using our 400 utterance in-set speaker evaluation set, our system's speaker recognition error rate was 0.25% over our closed set of 35 enrolled speakers (i.e., only one misrecognition in 400 trials). Thus we have achieved roughly the same error rate as on YOHO, but only with a much smaller set of speakers.

4. SUMMARY AND FUTURE WORK

In summary, our initial study in biometric fusion for user verification has demonstrated the benefits of combining face and speaker identification even when one of the biometric techniques has superior performance to the other. A 90% reduction in user verification equal error rate was observed when our speaker identification system was fused with a face identification system. This result was achieved with a system that forces the user to provide a frontal image that can be automatically detected with a high-level of confidence. By adjusting the confidence-level of the face detector, the system can reduce the inconvenience of re-capturing images when the face detector fails, but at the expense of reduced user verification accuracy.

Though this study demonstrated the feasibility of our approach, our current evaluation set is quite small. In future work we plan to expand the size of evaluation set and examine the specific types of errors the system makes. We also plan to investigate the performance of the system under the conditions where impostors are specifically selected based on resemblances of their voice or facial properties (i.e., same gender or ethnicity) to particular enrolled users.

5. ACKNOWLEDGMENTS

The authors wish to thank Dave Dopson and Ken Steele, who helped in the development of the application; and Jane Wu who assisted with data evaluation. This work was supported in part by the MIT Oxygen Alliance.

6. REFERENCES

[1] T. Hazen, E. Weinstein, and A. Park, "Towards robust person recognition on handheld devices using face and speaker identification technologies," in *Proc. of Int.*

Conf. on Multimodal Interfaces, Vancouver, Canada, November 2003.

[2] A. Park and T. Hazen, "ASR dependent techniques for speaker identification," in *Proc. of Int. Conf. on Spoken Language Processing*, Denver, Colorado, September 2002, pp. 1337–1340.

[3] B. Heisele, P. Ho, and T. Poggio, "Face recognition with support vector machines: Global versus component-based approach," in *Proc. of Int. Conf. on Computer Vision*, Vancouver, Canada, July 2001, vol. 2, pp. 688–694.

[4] B. Heisele, T. Serre, S. Prentice, and T. Poggio., "Hierarchical classification and feature reduction for fast face detection with support vector machines," *Pattern Recognition*, vol. 36, pp. 2007–2017, 2003.

[5] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, Germany, 1995.

[6] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Halmstad, Sweden, June 2001, pp. 348–353.

[7] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, October 1995.

[8] J. Kittler, Y. Li, J. Matas, and M. Sanchez, "Combining evidence in multimodal personal identity recognition systems," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Crans-Montana, Switzerland, March 1997, pp. 327–334.

[9] E. Bigün, J. Bigün, B. Duc, and S. Fischer, "Expert conciliation for multi modal person authentication systems by Bayesian statistics," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Crans-Montana, Switzerland, March 1997, pp. 291–300.

[10] A. Ross, A. Jain, and J. Qian, "Information fusion in biometrics," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentification*, Halmstad, Sweden, June 2001, pp. 354–359.

[11] E. Weinstein, P. Ho, B. Heisele, T. Poggio, K. Steele, and A. Agarwal, "Handheld face identification technology in a pervasive computing environment," in *Short Paper Proceedings, Pervasive 2002*, Zurich, Switzerland, August 2002, pp. 48–54.

[12] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001, pp. 511–518.

[13] J. Campbell, "Testing with the YOHO CD-ROM voice verification corpus," in *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, May 1998, pp. 341–344.

Invited Speaker

Josef Bigun
Halmstad University

Multimodal Biometric Authentication for Mobile Communication

Abstract

The elements of multi-modal authentication along with system models are presented. These include the machine experts as well as machine supervisors. These will be contrasted to human performance. In particular fingerprint and speech based systems will serve as illustrations of a mobile authentication application. A signal adaptive supervisor, based on the input biometric signal quality, will be discussed. Experimental results on data collected from mobile telephones are reported demonstrating the benefits of the proposed scheme in mobile communication systems. The presentation is based on these studies, for which the research documented in has been instrumental.

Non-Linear Variance Reduction Techniques in Biometric Authentication

Norman Poh and Samy Bengio
IDIAP, Rue du Simplon 4, Martigny, CH-1920 Martigny, Switzerland
Email: {norman,bengio}@idiap.ch
Telephone: (41) 27–721.77.53 Fax: (41) 27–721.77.12

Abstract— **In this paper, several approaches that can be used to improve biometric authentication applications are proposed. The idea is inspired by the ensemble approach, i.e., the use of several classifiers to solve a problem. Compared to using only one classifier, the ensemble of classifiers has the advantage of reducing the overall variance of the system. Instead of using multiple classifiers, we propose here to examine other possible means of variance reduction (VR), namely through the use of multiple synthetic samples, different extractors (features) and biometric modalities. The scores are combined using the average operator, Multi-Layer Perceptrons and Support Vector Machines. It is found empirically that VR via modalities is the best technique, followed by VR via extractors, VR via classifiers and VR via synthetic samples. This decreasing order of effectiveness is due to the corresponding degree of independence of the combined objects. The theoretical and empirical findings show that experts combined via VR techniques *always* perform better than the average of their participating experts. Furthermore, in practice, *most* combined experts perform better than any of their participating experts.**

I. Introduction

Biometric authentication (BA) is the problem of verifying an identity claim using a person's behavioural and physiological characteristics. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially "who one is", i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness. Examples of biometric modalities are fingerprints, faces, voice, hand-geometry and retina scans [1].

To date, biometric-based security systems (devices, algorithms, architectures) still have room for improvement, particularly in their accuracy, tolerance to various noisy environments and scalability as the number of individuals increases. The focus of this study is to improve system accuracy by directly minimising the effects of noise via various variance reduction techniques. Biometric data is often noisy because of deformable templates, corruption by environmental noise, variability over time and occlusion by the user's accessories. The higher the noise, the less reliable the biometric system becomes.

Advancements in biometrics show two emerging solutions: combining several biometric modalities [2], [3] (often called multi-modal biometrics) and combining several samples of a single biometric modality [4]. These techniques are related to *variance reduction* (VR). This is a phenomenon originating from combining classifier scores. Specifically, by combining the outputs of N classifier scores using an average operator (in the simplest case), one can reduce the variance of the combined score, with respect to the target score, by a factor of N [5, Chap. 9], if the classifier scores are not correlated (or independent from each another). On the other hand, in the extreme case, when they are completely correlated (dependent on each other), there will be no reduction in variance at all [6].

In the context of BA, when one combines several biometric modalities or several samples, one indeed exploits the independence of each modality and sample, respectively. In this work, we examine several other ways to exploit this (often partial) independence, namely via extractors, classifiers and synthetic samples. In short, all these methods can be termed as follows: Variance Reduction (VR) via classifiers, VR via extractors, VR via samples and VR via (biometric) modalities.

In our opinion, VR techniques have the potential to improve the accuracy of BA systems because better classifiers or ensemble methods, feature extraction algorithms and biometric-enabled sensors are emerging. Instead of choosing one best technique (best features, classifiers, etc), VR techniques propose to combine these new algorithms with existing techniques (features, classifiers) to obtain improved results, whenever this is feasible. The added overhead cost will be computation time and possibly hardware cost in the case of adding new sensors (as opposed to other VR techniques which *do not require* any extra hardware).

II. Variance Reduction in Biometric Authentication

A. Variance Reduction

This section presents a brief findings on the theory of variance reduction (VR). Details can be found in [6].

A person requesting an access can be measured by his or her biometric data. Let this biometric data be $\mathbf{x}$. This measurement can be done by several methods, to be explored later. Let i denote the i-th extract of $\mathbf{x}$ by a given method. For the sake of comprehension, one method to do so is to use multiple samples. Thus, in this case, i denotes the i-th sample. If the chosen method uses multiple biometric modalities, then i refers to the i-th biometric modality. Let the measured relationship be denoted as $y_i(\mathbf{x})$. It can be thought as the i-th response (of the sample or modality, for instance) given by a biometric system. Typically, this output (e.g. score) is used to make the accept/reject decision. $y_i(\mathbf{x})$ can be decomposed into two components, as follows:

$$y_i(\mathbf{x}) = h(\mathbf{x}) + \eta_i(\mathbf{x}), \quad (1)$$

where $h(\mathbf{x})$ is the "target" function that one wishes to estimate and $\eta_i(\mathbf{x})$ is a random additive noise with zero mean, also dependent on $\mathbf{x}$.

Let N be the number of trials, (e.g., the number of samples, assuming that the chosen method uses multiple samples hereinafter).

The mean of y over N trials, denoted as $\bar{y}(\mathbf{x})$ is:

$$\bar{y}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} y_i(\mathbf{x}). \tag{2}$$

When N samples are available and they are used separately, the *average of variance* made by each sample, independently, is:

$$\text{VAR}_{AV}(\mathbf{x}) \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\text{VAR}[y_i(\mathbf{x})], \tag{3}$$

where VAR[x] is the variance of x.

The variance as a result of averaging (or *variance of average*) due to Eqn. (2) is defined as:

$$\text{VAR}_{COM}(\mathbf{x}) \quad = \quad E[(\bar{y}(\mathbf{x}) - h(\mathbf{x}))^2], \tag{4}$$

where E[x] is the expectation of x. In our previous work [6], it has been shown that:

$$\frac{1}{N}\text{VAR}_{AV}(\mathbf{x}) \leq \text{VAR}_{COM}(\mathbf{x}) \leq \text{VAR}_{AV}(\mathbf{x}). \tag{5}$$

This equation shows that when scores $y_i, i = 1, \ldots, N$ are uncorrelated, the variance of average is reduced by a factor of $1/N$ with respect to the average of variance. On the other hand, when the scores are totally correlated, there is no reduction of variance, with respect to the average of variance.

To measure *explicitly* the factor of reduction, we introduce α, which can be defined as follows:

$$\alpha = \frac{\text{VAR}_{AV}(\mathbf{x})}{\text{VAR}_{COM}(\mathbf{x})}. \tag{6}$$

By dividing Eqn (5) by VAR$_{COM}$ and rearranging it, we can deduce that $1 \leq \alpha \leq N$.

B. Variance Reduction and Classification Reduction

Fig. 1 illustrates the effect of averaging scores in a two-class problem, such as in BA where an identity claim could belong either to a client or an impostor. Let us assume that the genuine user scores in a situation where 3 samples are available but are used separately, follow a normal distribution of mean 1.0 and variance (VAR$_{AV}(\mathbf{x})$ of genuine users) 0.9, denoted as $\mathcal{N}(1, \sqrt{0.9})$, and that the impostor scores (in the mentioned situation) follow a normal distribution of $\mathcal{N}(-1, \sqrt{0.6})$ (both graphs are plotted with "+"). If for each access, the 3 scores are used, according to Equation 6, the variance of the resulting distribution will be reduced by a factor (which is the value α defined in Equation 6) of 3 or less. Both resulting distributions are plotted with "o". Note the area where both the distributions cross before and after. The later area is shaded in Fig. 1. This area corresponds to the zone where minimum amount of mistakes will be committed given that the threshold is optimal[1]. Decreasing this area implies an improvement in the performance of the system.

C. Variance Reduction and Correlation in Input Score Space

From the previous section, it is obvious that by reducing the variance, the classification results should be improved. How much variance can be reduced depends on how correlated the input scores are. The correlation between scores of two experts can be examined by plotting their scores on a 2D-plan, with one axis for each expert. This is shown in Figs. 2 and 3. The first figure shows a scatter-plot of scores taken from two experts working on the *same* features. The second figure shows a scatter-plot of scores taken from two experts working on *different biometric modalities*. Details of the experts are explained in Sec. IV. As can be seen, the scores of the former overlaps more than the latter, i.e., if a boundary is to be drawn between clients and impostors scores, it would be more difficult for the former problem than the latter problem. Note that overlapping occurs when both experts make the same errors. Thus, there will be more classification errors in the former problem than in the latter.

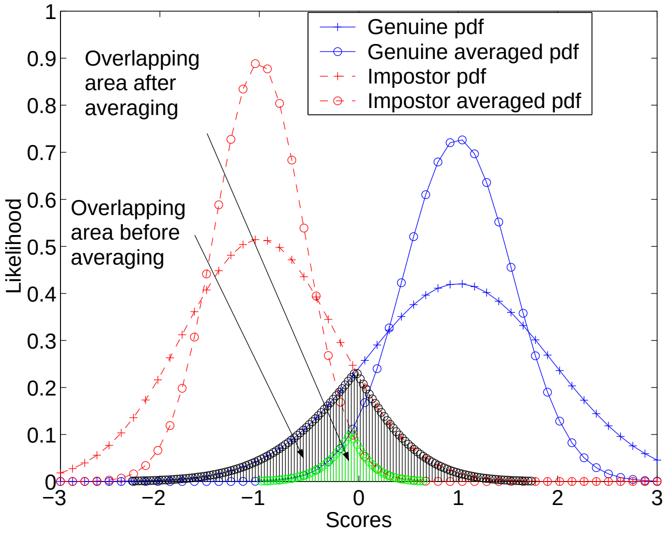


Fig. 1. Averaging score distributions in a two-class problem

D. Exploring Various Variance Reduction Techniques

This section explores various variance reduction (VR) techniques that can be applied to the BA problem. A BA system can be viewed as a system consisting of sensors, extractors, classifiers and a supervisor. Sensors such as cameras are responsible to capture a person's biometric traits. Extractors are responsible for extracting salient features that are useful for discriminating a person from others. Classifiers (also referred to as "experts") are responsible for comparing the extracted features to previously stored features that are known to belong to the person. Finally, in the context of multiple modalities, features, classifiers or samples, a supervisor is needed to merge all the results. A survey of different fusion techniques can be found in [7].

This serial concatenation process of sensors, extractors, classifiers and a supervisor shows that errors may accumulate along the chain because each module depends on the previous module. An important finding in Sec. II-A [6] is that it is beneficial to increase the number of processes. For instance, one can use more samples or more biometric modalities. In these two cases, N will be the number of samples and modalities, respectively. This is because by increasing N, one can decrease the variance further, regardless of how correlated the scores obtained from these N experts are. Note that in the work of Kittler *et al* [4], they showed that by increasing N samples up to a limit, there is no more gain in accuracy. When this happens, they deem the system to be "saturated". In our context, we expand N through different methods, as follows:

- **Multiple Biometric Modalities**. Each modality has its own feature set and classifiers. In other words, they operate independently of each other [7]–[10]
- **Multiple Samples**. Samples could be real [4] or virtually generated [11].

[1] Optimal in the Bayes sense, when (1) the cost and (2) probability of both types of errors (i.e., false acceptances and false rejections) are equal.

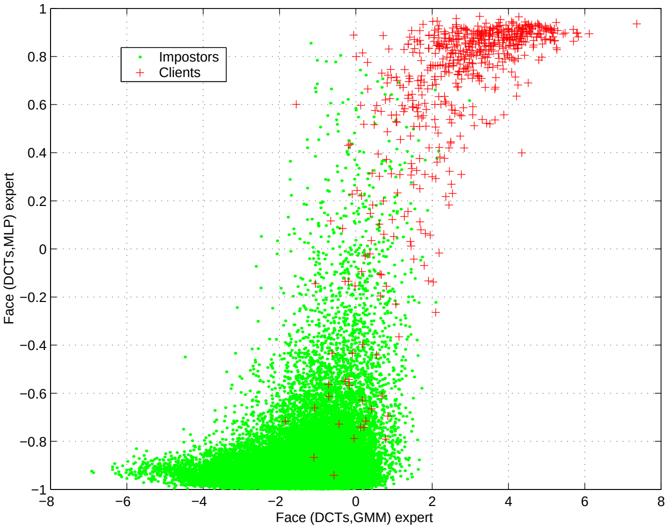


Fig. 2. Scores from experts of different features

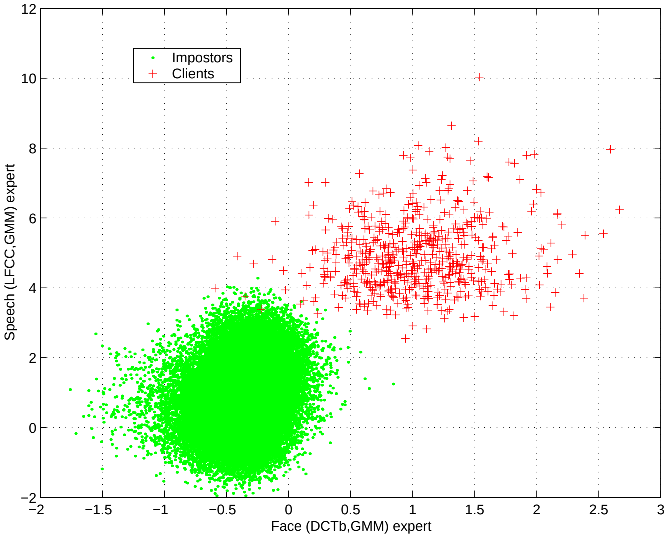


Fig. 3. Scores from experts of different biometric modalities

- **Multiple Extractors**. Each feature is classified by a classifier independently of other features [12]–[14].
- **Multiple Classifiers**. All classifiers receive the same input features. Heterogeneous types of classifiers can be used. Unstable homegenous classifiers such as Multi-Layer Perceptrons (MLPs) trained by bagging or with different hidden units can also be used. In general, many ensemble methods such as bagging, boosting, via Error-Correcting Output-Coding fall in this category [15], [16].

For each method mentioned above, the problem now is to combine these N scores. This is treated in the next subsection.

E. Fusions in Variance Reduction Techniques

In Sec. II-A, it has been illustrated that correlation of scores in the input space plays a vital role in determining the success of the resultant combined system. Furthermore, by simple averaging of N scores, it has been shown that the variance of the resultant combined score can be reduced by a factor between 1 and N with respect to the average of variance.

Instead of using simple averaging, one could have used weighted average, or even non-linear techniques such as MLPs and Support Vector Machines (SVMs) [5]. In the latter two cases however, one needs to select carefully the various hyper-parameters of these models (such as the number of hidden units in the MLPs or the kernel parameters in the SVMs). According to the Statistical Learning Theory [17], the expected performance of a model such as an MLP or an SVM on new data depends on the *capacity* of the set of functions the model can approximate. If the capacity is too small, the desired function might not be in the set of functions, while if it too high, several apparently good functions could be approximated, with the risk of selecting a bad one. This phenomenon is often called *over-training*. Although this capacity cannot unfortunately be explicitely estimated for complex set of functions such as MLPs and SVMs, its ordering can be used to select efficiently the corresponding hyper-parameters using some sort of validation technique. One such method is the K-fold cross-validation.

Algorithm 1 shows how K-fold cross-validation can be used to estimate the correct value of the hyper-parameters of our fusion model, as well as the decision threshold used in the case of authentication. The basic framework of the algorithm is as follows: first perform K-fold cross-validation on the training set by varying the capacity parameter, and for each capacity parameter, select the corresponding decision threshold that minimises Half Total Error Rate (HTER)[2]; then choose the best hyper-parameter according to this criterion and perform normal training with the best hyper-parameter on the whole training set; finally test the resultant classifier on the test set [8] with HTER evaluated on the previously found decision threshold.

There are several points to note concerning Algorithm 1: $\mathcal{Z}$ is a set of labelled examples of the form $(\mathcal{X}, \mathcal{Y})$, where the first term is a set of patterns and the second term is a set of corresponding labels. The "train" function receives a hyper-parameter θ and a training set, and outputs an optimal classifier $\hat{F}$ by minimising the HTER on the training set. The "test" function receives a classifier $\hat{F}$ and a set of examples, and outputs a set of scores for each associated example. The "thrd$_{HTER}$" function returns a *decision threshold* that minimises HTER by minimising $|\text{FAR}(\Delta) - \text{FRR}(\Delta)|$ with respect to the threshold Δ (FAR(Δ) and FRR(Δ) are false acceptance and false rejection rates, as a function of Δ) while L_{HTER} returns the HTER *value* for a particular decision threshold. What makes this cross-validation different from classical cross-validation is that there is only one single decision threshold and the corresponding HTER value for all the held-out folds and for a given hyper-parameter θ. This is because it is logical to union scores of all held-out folds into one single set of scores to select the decision threshold (and obtain the corresponding HTER).

F. Fusions For VR via Samples

All the VR techniques discussed earlier can be treated in a general manner, except VR via samples. This is because the ordering of scores induced by samples are not important. Simply concatenating the scores and feeding them to a classifier may not be an optimal solution. Another problem that might arise is that when there are many scores, possibly in the range of hundreds (one can generate as many virtual scores as one wishes), matching should be done in terms of their distribution instead. We hence propose two solutions to handle this: 1) estimate the likelihood of the set of virtual scores when coming from either a client or an impostor distribution; 2) estimate the distribution of the scores so that matching will be performed between a competing

[2] HTER is defined as (FAR+FRR)/2, where FAR is False Acceptance Rate and FRR is False Rejection Rate.

Algorithm 1 Risk Estimation $(\Theta, K, \mathcal{Z}^{train}, \mathcal{Z}^{test})$

```
REM: Risk Estimation with K-fold Validation. See [8].
$\Theta$ : a set of values for a given hyper-parameter
$\mathcal{Z}^i$ : a tuple $(\mathcal{X}^i, \mathcal{Y}^i)$, for $i \in \{train, test\}$ where
$\mathcal{X}$ : a set of patterns. Each pattern contains scores/hypothesis from base experts
$\mathcal{Y}$ : a set of labels $\in \{client, impostor\}$
Let $\cup_{k=1}^{K} \mathcal{Z}^k = \mathcal{Z}^{train}$
for each hyper-parameter $\theta \in \Theta$ do
  for each $k = 1, \ldots, K$ do
    $\hat{F}_\theta$ = train($\theta$, $\cup_{j=1, j \neq k}^{K} \mathcal{Z}^j$)
    $\hat{\mathcal{Y}}_\theta^k$ = test($\hat{F}_\theta$, $\mathcal{X}^k$)
  end for
  $\Delta_\theta = \text{thrd}_{HTER} \left( \{\hat{\mathcal{Y}}_\theta^k\}_{k=1}^K, \{\mathcal{Y}^k\}_{k=1}^K \right)$
end for
$\theta^* = \arg\min_\theta \left( L_{HTER} \left( \Delta_\theta, \{\hat{\mathcal{Y}}_\theta^k\}_{k=1}^K, \{\mathcal{Y}^k\}_{k=1}^K \right) \right)$
$\hat{F}_{\theta^*}$ = train($\theta^*$, $\mathcal{Z}^{train}$)
$\hat{\mathcal{Y}}_{\theta^*}^{test}$ = test($\hat{F}_{\theta^*}$, $\mathcal{X}^{test}$)
return $L_{HTER}(\Delta_{\theta^*}, \hat{\mathcal{Y}}_{\theta^*}^{test}, \mathcal{Y}^{test})$
```

client and an impostor distribution. Both approaches assume that the scores are generated independently from some unknown distributions. Of course this independence assumption is not true, but it is good enough for most practical problems.

The first approach is carried out using Gaussian Mixture Models (GMMs) to model the scores. First estimate the client and impostor distributions using GMMs by separately maximising the likelihood of the client and impostor scores using the Expectation-Maximisation algorithm [5]. During an access request with one real biometric sample, a set of synthetic samples and hence a set of scores are generated. These scores will be fed to the client and an impostor GMM score distribution. Let $\log p(\mathbf{x}|\theta_C)$ be the log likelihood of the set of scores $\mathbf{x}$ given the client GMM model θ_C and $\log p(\mathbf{x}|\theta_I)$ be the same term but for the impostor model. The decision is often taken using the so-called log-likelihood ratio:

$$s = \log p(\mathbf{x}|\theta_C) - \log p(\mathbf{x}|\theta_I)$$

In the second approach, we propose to first model the distribution of these synthetic scores using a Parzen window non parametric density model [5, Chap. 2] and then compute the relative entropy of each distribution, which is defined as follows:

$$L(p, q) = -\sum_i p(y_i) \log \frac{q(y_i)}{p(y_i)}, \qquad (7)$$

where q and p are two distributions. Entropy can be regarded as a distortion of $q(y)$ from $p(y)$. This alone does not give discriminative information. To do so, entropies of a client and an impostor distribution should be used together. Let $L(p_C, q)$ be the entropy of $q(y)$ with respect to a client distribution and $L(p_I, q)$ be that of $q(y)$ with respect to an impostor distribution. Then the difference between these two entropies, can be defined as:

$$s = L(p_I, q) - L(p_C, q).$$

When $s > 0$, the distortion of $q(y)$ from an impostor distribution is greater than that of a client distribution, which reflects how likely a set of synthetic scores belong to a client. In fact, for both approaches, $s > \triangle$ is used instead, where $\triangle$ is a threshold chosen *a priori* according to the HTER criterion.

III. Experimental Settings

A. XM2VTS Database Description

The XM2VTS database [18] contains synchronised video and speech data from 295 subjects, recorded during four sessions taken at one month intervals. On each session, two recordings were made, each consisting of a speech shot and a head shot. The speech shot consisted of frontal face and speech recordings of each subject during the recital of a sentence.

The database is 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 (Eval) was used to compute the decision thresholds (as well as other hyper-parameters) used by classifiers. Finally, the test set (Test) was used to estimate the performance.

The 295 subjects were divided into a set of 200 clients, 25 evaluation impostors and 70 test impostors. There exists two configurations or two different partitioning approaches of the training and evaluation sets. They are called Lausanne Protocol I and II, denoted as **LP1** and **LP2** in this paper. Thus, besides the data for training the model, the following four data sets are available for evaluating the performance: LP1 Eval, LP1 Test, LP2 Eval and LP2 Test. Note that LP1 Eval and LP2 Eval are used to calculate the optimal thresholds that will be used in LP1 Test and LP2 Test, respectively. Results are reported only for the test sets, in order to be as unbiased as possible (using an *a priori* selected threshold). Table I is the summary of the data. In both configurations, the test set remains the same. However, there are three training shots per client for LP1 and four training shots per client for LP2. More details can be found in [19].

B. Feature Extraction

For the face data, a bounding box is placed on a face according to manually located eye co-ordinates. This assumes a perfect face detection[3]. The face is cropped and the extracted sub-image is downsized to a 40×30 (rows $\times$ columns) image. After enhancement and smoothing, the face image is represented as a feature vector with a dimensionality of 1200.

In addition to these normalised features, RGB (Red-Green-Blue) histogram features are used. For each colour channel, a histogram is built using 32 discrete bins. Hence, the histograms of three channels, when concatenated, form a feature vector of 96 elements. More details about this method, including experiments, can be obtained from [20].

Another feature set derived from Discrete Cosine Transform (DCT) coefficients [21], [22] has also given good performance. The idea is

TABLE I
The Lausanne Protocols of XM2VTS database

Data sets	Lausanne Protocols	
	LP1	LP2
Training client accesses	3	4
Evaluation client accesses	600 (3 × 200)	400 (2 × 200)
Evaluation impostor accesses	40,000 (25 × 8 × 200)	
Test client accesses	400 (2 × 200)	
Test impostor accesses	112,000 (70 × 8 × 200)	

[3] Hence, even if this is often done in the literature, the final results using face scores could be optimistically biased due to this manual detection step. Note on the other hand that due to the clean and controlled quality of XM2VTS, automatic detectors often yield detection rates of around 99%.

to divide images into overlapping blocks. For each block, a subset of DCT coefficients is computed. The horizontal and vertical deltas of several DCT coefficients are also found. It has been shown that this feature set (referred to as DCTmod2) has better performance than features derived from Principal Component Analysis [21].

For the speech data, the feature sets used in the experiments are Linear Filter-bank Cepstral Coefficients (LFCC) [23], Phase Auto-correlation derived Mel-scale Frequency Cepstrum Coefficients (PAC) [24] and Mean-Subtracted Spectral Subband Centroids (SSC) [25]. The speech/silence segmentation is done using two competing Gaussians trained in an unsupervised way by maximising the likelihood of the data given a mixture of the 2 Gaussians. One Gaussian will end up modelling the speech and the other will end up modelling the non-speech feature frames [26]. In general, the segmentation given by this technique is satisfactory.

IV. Results

In order to analyse the effects due to VR techniques, we first present the baseline experimental results. This is followed by results obtained by various VR techniques. Note that all results reported here are in terms of **percentage of HTER**, the thresholds are all selected **a priori** (i.e., tuned on the training set, hence the threshold is *completely independent* of the test set and is thus unbiased), and for the combination strategy, **only two experts are used** each time.

A. Baseline Performance on The XM2VTS Database

The face baseline experts are based on the following features:

1) **FH**: normalised **f**ace image concatenated with its RGB **H**istogram (thus the abbreviation **FH**)
2) **DCTs**: DCTmod2 features extracted from face images with a size of 40×32 (rows × columns) pixels. The DCT coefficients are calculated from an 8 × 8 window with horizontal and vertical overlaps of 50%, i.e., 4 pixels in each direction. Neighbouring windows are used to calculate the “delta” features. The result is a set of 35 feature vectors, each having a dimensionality of 18. (**s** indicates the use of this small image compared to the bigger size image with the abbreviation **b**.)
3) **DCTb**: Similar to DCTs except that the input face image has 80×64 pixels. The result is a set of 221 feature vectors, each having a dimensionality of 18.

The speech baseline experts are based on the following features:

1) **LFCC**: The Linear Filter-bank Cepstral Coefficient (LFCC) speech features were computed with 24 linearly-spaced filters on each frame of Fourier coefficients sampled with a window length of 20 milliseconds and each window moved at a rate of 10 milliseconds. 16 DCT coefficients are computed to decorrelate the 24 coefficients (log of power spectrum) obtained from the linear filter-bank. The first temporal derivatives are added to the feature set.
2) **PAC**: The PAC-MFCC features are derived with a window length of 20 miliseconds and each window moves at a rate of 10 miliseconds. 20 DCT coefficients are computed to decorrelate the 30 coefficients obtained from the Mel-scale filter-bank. The first temporal derivatives are added to the feature set.
3) **SSC**: The mean-subtracted SSCs are obtained from 16 coefficients. The γ parameter, which is a parameter that raises the power spectrum and controls how much influence the centroid, is set to 0.7. Also The first temporal derivatives are added to the feature set.

Two different types of classifiers were used for these experiments: an MLP and a Bayes Classifier using GMMs to estimate the class distributions [5]. While in theory both classifiers could be trained using any of the previously defined feature sets, in practice only some specific combinations appear to yield reasonable performance.

Whatever the classifier is, the hyper-parameters (e.g. the number of hidden units for MLPs or the number of Gaussian components for GMMs) are tuned on the evaluation set LP1 Eval. The same set of hyper-parameters are used in both LP1 and LP2 configurations of the XM2VTS database.

For each client-specific MLP, the samples associated to the client are treated as positive patterns while all other samples *not* associated to the client are treated as negative patterns. All MLPs reported here were trained using the stochastic version of the error-backpropagation training algorithm [5].

For the GMMs, two competing models are often needed: a world and a client-dependent model. Initially, a world model is first trained from an external database (or a sufficiently large data set) using the standard Expectation-Maximisation algorithm [5]. The world model is then adapted for each client to the corresponding client data using the Maximum-A-Posteriori adaptation [27] algorithm.

The baseline experiments based on DCTmod2 feature extraction were reported in [22] while those based on normalised face images and RGB histograms (FH features) were reported in [20]. Details of the experiments, coded in the pair **(feature, classifier)**, for the face experts, are as follows:

1) **(FH, MLP)** Features are normalised **F**ace concatenated with **H**istogram features. The client-dependent classifier used is an MLP with 20 hidden units. The MLP is trained with geometrically transformed images [20].
2) **(DCTs, GMM)** The face features are the DCTmod2 features calculated from an input face image of 40×32 pixels, hence, resulting in a sequence of 35 feature vectors each having 18 dimensions. There are 64 Gaussian components in the GMM. The world model is trained using *all the clients* in the training set [22].
3) **(DCTb, GMM)** Similar to (DCTs,GMM), except that the features used are DCTmod2 features calculated from an input face image of 80×64 pixels. This produces in a sequence of 221 feature vectors each having 18 dimensions. The corresponding GMM has 512 Gaussian components [22].
4) **(DCTs, MLP)** Features are the same as those in (DCTs,GMM) except that an MLP is used in place of a GMM. The MLP has 32 hidden units [22]. Note that in this case a training example consists of a *big single* feature vector with a dimensionality of 35×18. This is done by simply concatenating 35 feature vectors each having 18 dimensions[4].
5) **(DCTb, MLP)** The features are the same as those in (DCTb,GMM) except that an MLP with 128 hidden units is used. Note that in this case the MLP in trained on a *single* feature vector with a dimensionality of 221×18 [22].

and for the speech experts:

1) **(LFCC, GMM)** This is the Linear Filter-bank Cepstral Coefficients (LFCC) obtained from the speech data of the XM2VTS database. The GMM has 200 Gaussian components, with the minimum relative variance of each Gaussian fixed to 0.5, and the MAP adaptation weight equals 0.1. This is the best known model currently available.
2) **(PAC, GMM)** The same GMM configuration as in LFCC is used. Note that in general, 200-300 Gaussian components

[4] This may explain why MLP, an inherently discriminative classifier, has worse performance compared to GMM, a generative classifier. With high dimensionality yet having only a few training examples, the MLP cannot be trained optimally. This may affect its generalisation on unseen examples. By treating the features as a sequence, GMM was able to generalise better and hence is more adapted to this feature set.

TABLE II

BASELINE PERFORMANCE IN HTER(%) OF DIFFERENT MODALITIES EVALUATED ON XM2VTS BASED ON *a priori* SELECTED THRESHOLDS

Data sets	(Features, classifiers)	FAR	FRR	HTER
Face LP1 Test	(FH,MLP)	1.751	2.000	1.875
Face LP1 Test	(DCTs,GMM)	4.454	4.000	4.227
Face LP1 Test	(DCTb,GMM)	1.840	1.500	1.670
Face LP1 Test	(DCTs,MLP)	3.219	3.500	3.359
Face LP1 Test	(DCTb,MLP)	4.443	8.000	6.221
Speech LP1 Test	(LFCC,GMM)	1.029	1.250	1.139
Speech LP1 Test	(PAC,GMM)	4.608	8.000	6.304
Speech LP1 Test	(SSC,GMM)	2.374	2.500	2.437
Face LP2 Test	(FH,MLP)	1.469	2.250	1.860
Face LP2 Test	(DCTb,GMM)	1.039	0.250	0.644
SpeechLP2 Test	(LFCC,GMM)	1.349	1.250	1.300
Speech LP2 Test	(PAC,GMM)	5.283	8.000	6.642
Speech LP2 Test	(SSC,GMM)	2.276	1.750	2.013

would give about 1% of difference of HTER.

3) **(SSC, GMM)** The same GMM configuration as in LFCC is used.

The baseline performances are shown in Table II.

As can be seen, among the face experiments, (DCTb,GMM) performs the best across all configurations while (DCTb,MLP) performs the worst. In the speech experiments, LFCC features are the best features, followed by SSC and PAC, in decreasing order of accuracy. Regardless of strong or weak classifiers, as long as their correlation is weak, they can be used in the VR techniques.

B. VR via Different Modalities, Extractors, Classifiers

Table III shows the results combining scores of two modalities, two extractors and two classifiers (working on the same feature space). The second to last column shows the mean HTER of each of the two underlying experts while the last column shows the minimum HTER of the two experts. The three sub-columns under the heading "joint HTER" are the HTERs of the combined experts via the mean operator, MLP and SVM. Numbers in bold are the best HTER among the three fusion methods. A quick examination of this table reveals that all combined experts via modalities are better than the best underlying expert (compare min HTER with the scores in the joint HTER). However, the combined experts via extractors and classifiers, as shown in Table IV, are not always better than their participating experts.

C. VR via Virtual Samples

The experiments on VR via samples are presented differently than the rest because they cannot be evaluated using the mean HTER and min HTER. Instead, the combined experts are compared to the original baseline experts (compare the first row of Table V against the other rows). The two numbers in bold are the best fusion technique for LP1 and LP2 configurations, respectively. The Entropy and GMM approaches are discussed in Sec. II-F. The median technique refers to combining synthetic scores using the median operator which is known to be robust to outlier scores. We note that the best fusion techniques on these datasets are the entropy approach and the GMM approach for LP1 and LP2, respectively. For LP1, the entropy approach is *significantly better* with 90% confidence level than the mean operator according to the McNemar's Test[5] [28] (i.e., with a difference of 0.006 HTER% between the two approaches). For LP2, the GMM approach is *significantly better* than the mean operator with 99% confidence level. This shows that exploiting the distribution of scores *is better* than using the simple mean operator.

TABLE III

PERFORMANCE IN (%) OF HTER OF VR VIA MODALITIES ON XM2VTS BASED ON *a priori* SELECTED THRESHOLDS

(a) Face experts and (LFCC,GMM) expert

Data sets	Face,	Joint HTER			mean	min
	Experts	mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP)	0.399	**0.366**	0.381	1.507	1.139
LP1 Test	(DCTs,GMM)	**0.537**	0.576	0.613	2.683	1.139
LP1 Test	(DCTb,GMM)	0.520	0.483	**0.475**	1.405	1.139
LP1 Test	(DCTs,MLP)	0.591	0.611	**0.587**	2.249	1.139
LP1 Test	(DCTb,MLP)	0.497	0.489	**0.485**	3.680	1.139
LP2 Test	(FH,MLP)	0.151	**0.150**	0.389	1.580	1.300
LP2 Test	(DCTb,GMM)	0.147	**0.130**	0.252	0.972	0.644

(b) Face experts and (PAC,GMM) expert

Data sets	Face,	Joint HTER			mean	min
	Experts	mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP)	1.114	**0.856**	0.970	4.090	1.875
LP1 Test	(DCTs,GMM)	1.407	1.425	**1.402**	5.266	4.227
LP1 Test	(DCTb,GMM)	**0.899**	0.900	0.923	3.987	1.670
LP1 Test	(DCTs,MLP)	1.248	1.056	**1.009**	4.832	3.359
LP1 Test	(DCTb,MLP)	3.978	**2.455**	2.664	6.263	6.221
LP2 Test	(FH,MLP)	1.282	**0.765**	0.855	4.251	1.860
LP2 Test	(DCTb,GMM)	0.243	**0.222**	0.431	3.643	0.644

(c) Face experts and (SSC,GMM) expert

Data sets	Face,	Joint HTER			mean	min
	Experts	mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP)	0.972	0.786	**0.742**	2.156	1.875
LP1 Test	(DCTs,GMM)	**1.028**	1.175	1.213	3.332	2.437
LP1 Test	(DCTb,GMM)	0.756	**0.704**	0.742	2.053	1.670
LP1 Test	(DCTs,MLP)	1.167	**0.829**	0.850	2.898	2.437
LP1 Test	(DCTb,MLP)	2.986	1.176	**1.121**	4.329	2.437
LP2 Test	(FH,MLP)	0.901	**0.302**	0.404	1.937	1.860
LP2 Test	(DCTb,GMM)	**0.049**	0.162	0.383	1.329	0.644

D. Evaluation of Experiments

Let us define two measures of gain so as to draw a summary of the experiments carried out above, as below:

$$\beta_{mean} = \frac{\text{mean}_i(\text{HTER}_i)}{\text{HTER}_c}$$

$$\beta_{min} = \frac{\text{min}_i(\text{HTER}_i)}{\text{HTER}_c},$$

where β_{mean} and β_{min} measure how many times the HTER of the combined expert c is smaller than the mean and the min HTER of the underlying experts $i = 1, \ldots, N$. β_{mean} is designed to verify Eq. 6, which is somewhat akin to α. According to the theoretical analysis presented in Sec. II-A, $\alpha \geq 1$ should be satisfied. The β_{min}, on the other hand, is a more realistic criterion, i.e., one wishes to

[5] This is done by calculating $((n_{01} - n_{10})^2 - 1)/(n_{01} + n_{10}) > p$ where p is the inverse function of $\mathcal{X}^2$ distribution (with 1 degree of freedom) at a desired confidence interval (i.e., 90%), and n_{01} and n_{10} are the number of *different* mistakes done by the two systems on the *same* accesses

TABLE IV

Performance in (%) of HTER of VR via extractors and classifiers on XM2VTS based on *a priori* selected thresholds

Data sets	(Features, classifiers)	Joint HTER			mean	min
		mean	MLP	SVM	HTER	HTER
LP1 Test	(FH,MLP) (DCTs,GMM)	1.641	**1.379**	1.393	3.051	1.875
LP1 Test	(FH,MLP) (DCTb,GMM)	**1.123**	1.151	1.528	1.772	1.670
LP1 Test	(FH,MLP) (DCTs,MLP)	**1.475**	1.667	1.476	2.617	1.875
LP1 Test	(FH,MLP) (DCTb,MLP)	1.948	**1.933**	1.938	4.048	1.875
LP1 Test	(LFCC,GMM) (SSC,GMM)	1.296	1.444	**1.142**	1.788	1.139
LP1 Test	(PAC,GMM) (SSC,GMM)	3.594	2.954	**2.663**	4.370	2.437
LP2 Test	(FH,MLP) (DCTb,GMM)	0.896	0.670	**0.488**	1.252	0.644
LP2 Test	(LFCC,GMM) (SSC,GMM)	1.107	**1.034**	1.063	1.656	1.300
LP2 Test	(PAC,GMM) (SSC,GMM)	2.614	2.316	**2.125**	4.328	2.013
LP1 Test	(DCTs,GMM) (DCTs,MLP)	2.873	**2.486**	2.697	3.793	3.359
LP1 Test	(DCTb,GMM) (DCTb,MLP)	2.898	1.532	**1.471**	3.946	1.670

TABLE V

Performance in (%) of HTER of different combination methods of synthetic scores.

Method	HTER	
	LP1	LP2
Original	1.875	1.737
Mean	1.612	1.518
Median	1.667	1.547
GMM	1.709	**1.493**
Entropy	**1.606**	1.559

TABLE VI

Comparison of various VR techniques based on all experiments carried out using β_{mean}, β_{min} and β_{real}

(a) β_{mean} of all experiments

VR techniques	Table	No. of exp.	Joint HTER		
			mean	MLP	SVM
Modalities	III(a) (all)	21	5.559 ±5.879	5.390 ±3.287	4.164 ±1.474
	III(a) (LFCC)	7	5.680 ±2.683	5.843 ±2.744	4.375 ±1.482
	III(a) (PAC)	7	5.086 ±4.459	5.999 ±4.686	4.694 ±1.869
	III(a) (SSC)	7	5.910 ±9.365	4.326 ±2.128	3.422 ±0.733
Extractors	IV	9	1.604 ±0.269	1.719 ±0.313	1.842 ±0.420
Classifiers	IV	2	1.341 ±0.029	2.051 ±0.742	2.044 ±0.902
Synthetic samples	V	2	1.154 ±0.0002	MLP and SVM not used; see (b)	

(b) β_{real} of VR via synthetic samples

Methods	Gain ratio
Mean	1.154 ± 0.000178
Median	1.124 ± 0.000002
GMM	1.130 ± 0.002198
Global Entropy	1.141 ± 0.001422
Local Entropy	0.854 ± 0.000028

(c) β_{min} of all VR techniques except synthetic samples

VR techniques	Table	No. of exp.	Joint HTER		
			mean	MLP	SVM
Modalities	III(a)	21	3.043	3.109	2.459
Extractors	III(b)	9	1.009	1.067	1.120
Classifiers	III(c)	2	0.873	1.221	1.190

obtain better performance than the underlying experts, but there is no analytical proof that $\beta_{min} \geq 1$.

The β_{mean} for each experiment are shown in Table VI(a) for VR via modalities, extractors and classifiers, (b) for VR via synthetic samples and (c) for the gain ratio β_{min}. Note that VR via synthetic samples cannot be evaluated with the β_{min} criterion. It can only be compared to its original method (i.e., with real samples). This gain ratio can be defined as:

$$\beta_{real} = \frac{\text{HTER}_{real}}{\text{HTER}_{c}},$$

where $real$ is the expert that takes real samples and c is the expert that combines scores obtained from synthetic samples (in addition to the real sample).

Note that the β_{mean} for VR via modalites are sub-divided into 3 parts according to the 3 baseline speech experts (LFCC,GMM), (SSC,GMM) and (PAC,GMM) in a *significantly* decreasing order of accuracy. In such situations, the β_{mean} for these 3 baselines still have comparable range of values, which are bigger than other VR techniques. One possible conclusion is that regardless of the degree of accuracy of participating experts, as long as they are weakly correlated, high β_{mean} can be achieved. Although the mean operator seems to perform the best in the overall VR via modalities based on β_{mean}, it should be noted that out of the 27 experiments in Table III, 4 experiments are best combined with the mean operator, while there are 10 and 7 best results for MLPs and SVMs, respectively. Moreover, the standard deviation of the mean operator is much larger than that of MLPs and SVMs. In these experiments, MLP turns out to be a good candidate for fusion in most situations for VR via modalities. It should be emphasized that the success application of MLPs or SVMs in this fusion problem depends largely on the correct capacity estimate of cross-validation.

Note that Table VI(a) shows that $\beta_{mean} \geq 1$ for all fusion techniques but in (c), $\beta_{min} \geq 1$ is only true for MLPs and SVMs, but not for the mean operator, which we cannot guarantee. According to β_{mean} *on the mean operator*, VR via modalities achieves the highest gain, followed by VR via extractors, VR via classifiers and VR via synthetic samples. A similar trend is observed in (c) according to β_{min}. Such ordering is not a coincidence. It reveals that the correlation is greater and greater in the list just mentioned. In other words, β_{mean} is inversely proportional to the correlation of the underlying experts. However, with MLP and SVM as non-linear fusion techniques, this ordering is slightly perturbed because both the β_{mean} and β_{min} show that VR via classifiers are *better* than VR via extractors. Clearly, in highly correlated problems such as these, non-linear fusion techniques are better than the simple mean operator (but they come at an increase in complexity).

V. Conclusions

Variance reduction (VR) is an important technique to increase accuracy in regression and classification problems. In this study, several approaches are explored to improve Biometric Authentication systems, namely VR via modalities, VR via extractors, VR via classifiers and VR via synthetic samples. The experiments carried out on the XM2VTS database show that the combined experts due to VR techniques *always* perform better than the average of their participating experts, which can be explained by VR using the mean operator. Furthermore, all combined experts via modalities outperform the best participating expert based on the HTER. By means of non-linear variance reduction techniques, i.e., the use of MLPs and SVMs for combing scores obtained from participating experts, empirical study shows that, in average, these techniques could produce better results than their participating experts, in the context of VR via extractors and classifiers. In the context of VR via samples, exploiting the distribution of synthetic scores using GMM or Parzen-windows is better than the mean operator. In short, this study shows that non-linear fusion techniques using MLPs and SVMs, and incorporating other *a priori* information (i.e., distribution of synthetic scores in the case of synthetic samples) are vital to achieve high gain of fusion. In highly correlated situations (i.e., VR via extractors and classifiers), non-linear fusion techniques are very useful. In weakly correalted situations (i.e., VR via modalities), the mean operator could be feasible but non-linear fusion techniques are still useful if the capacity search using cross-validation is reliable. As new and more powerful extraction and classification algorithms become available, they can all be integrated into the VR framework. Therefore, VR techniques are potentially very useful for biometric authentication.

Acknowledgement

The authors thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research on ”Interactive Multimodal Information Management (IM2)” and Conrad Sanderson for his comments and suggestions.

References

[1] A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Person Identification in a Networked Society*. Kluwer Publications, 1999.

[2] L. Hong, A. Jain, and S. Pankanti, “Can Multibiometrics Improve Performance?” Computer Science and Engineering, Michigan State University, East Lansing, Michigan, Tech. Rep. MSU-CSE-99-39, 1999.

[3] N. Poh and J. Korczak, “Hybrid Biometric Authentication System Using Face and Voice Features,” in *3rd Int’l. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA’01)*, Halmstad, 2001, pp. 348–353.

[4] J. Kittler, G. Matas, K. Jonsson, and M. Sanchez, “Combining Evidence in Personal Identity Verification Systems,” *Pattern Recognition Letters*, vol. 18, no. 9, pp. 845–852, 1997.

[5] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1999.

[6] N. Poh and S. Bengio, “Variance Reduction Techniques in Biometric Authentication,” IDIAP, Martigny, Switzerland, Researh Report 03-17, 2003.

[7] C. Sanderson and K. K. Paliwal, “Information Fusion and Person Verification Using Speech & Face Information,” IDIAP, Martigny, Research Report 02-33, 2002.

[8] S. Bengio, C. Marcel, S. Marcel, and J. Mariéthoz, “Confidence Measures for Multimodal Identity Verification,” *Information Fusion*, vol. 3, no. 4, pp. 267–276, 2002.

[9] B. Duc, E. S. Bigun, J. Bigun, G. Maitre, and S. Fischer, “Fusion of Audio and Video Information for Multi Modal Person Authentication,” *Pattern Recognition Letters*, vol. 18, pp. 835–843, 1997.

[10] L. Hong and A. Jain, “Multi-Model Biometrics,” in *Biometrics: Person Identification in Networked Society*, 1999.

[11] N. Poh, S. Marcel, and S. Bengio, “Improving Face Authetication Using Virtual Samples,” in *IEEE Int’l Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, 2003, pp. 233–236 (Vol. 3).

[12] R. Brunelli and D. Falavigna, “Personal Identification Using Multiple Cues,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, 1995.

[13] F. Smeraldi, N. Capdevielle, and J. Bigun, “Face Authentication by Retinotopic Sampling of the Gabor Decomposition and Support Vector Machines,” in *Proc. 2nd Int’l Conf. Audio and Video Based Biometric Person Authentication (AVBPA’99)*, Washington DC, 1999, pp. 125–129.

[14] J. Luettin, “Visual Speech and Speaker Recognition,” Ph.D. dissertation, Department of Computer Science, University of Sheffield, 1997.

[15] T. Dietterich, “Ensemble Methods in Machine Learning,” in *Multiple Classifier Systems*, 2000, pp. 1–15.

[16] L. Breiman, “Bagging Predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[17] V. N. Vapnik, *Statistical Learning Theory*. Springer, 1998.

[18] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, “Comparison of Face Verification Results on the XM2VTS Database,” in *Proc. 15th Int’l Conf. Pattern Recognition*, Barcelona, 2000, pp. 858–863 (Vol. 4).

[19] J. Lüttin, “Evaluation Protocol for the XM2FDB Database (Lausanne Protocol),” IDIAP, Martigny, Switzerland, Communication 98-05, 1998.

[20] S. Marcel and S. Bengio, “Improving Face Verification Using Skin Color Information,” in *Proc. 16th Int. Conf. on Pattern Recognition*, Quebec, 2002.

[21] C. Sanderson and K. Paliwal, “Fast Features for Face Authentication Under Illumination Direction Changes,” *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2409–2419, 2003.

[22] F. Cardinaux, C. Sanderson, and S. Marcel, “Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS,” in *4th Int’l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA’03)*, Guildford, 2003, pp. 911–920.

[23] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Oxford University Press, 1993.

[24] S. Ikbal, H. Misra, and H. Bourlard, “Phase Auto-Correlation (PAC) derived Robust Speech Features,” in *Proc. IEEE Int’l Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, Hong Kong, 2003.

[25] K. K. Paliwal, “Spectral Subband Centroids Features for Speech Recognition,” in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, 1998, pp. 617–620 (Vol. 2).

[26] J. Mariéthoz and S. Bengio, “A Comparative Study of Adaptation Methods for Speaker Verification,” in *Int’l Conf. Spoken Language Processing (ICSLP)*, Denver, 2002, pp. 581–584.

[27] J. Gauvain and C.-H. Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Obervation of Markov Chains,” *IEEE Tran. Speech Audio Processing*, vol. 2, pp. 290–298, 1994.

[28] T. G. Dietterich, “Approximate Statistical Test for Comparing Supervised Classification Learning Algorithms,” *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

Temporal Integration for Continuous Multimodal Biometrics

Alphan Altinok and Matthew Turk
Computer Science Department
University of California, Santa Barbara
Santa Barbara, California 93106
$\{$*alphan, mturk*$\}$*@cs.ucsb.edu*

Abstract

Typically, biometric systems authenticate the user at a particular moment in time, granting or denying access to resources for the complete session. This model of authentication does not appropriately address environments where a different individual may take over a system from the original user (either willingly or otherwise). We propose a multimodal system that performs authentication continuously *by integrating information temporally as well as across modalities. Such continuous authentication provides ongoing (rather than one-time) verification and can easily be coupled with another system for dynamically adjusting access to privileges accordingly.*

We present an initial approach for temporal integration based on uncertainty propagation over time for estimating channel output distribution from recent history, and classification with uncertainty. Our method operates continuously by computing expected values as a function of time differences. Our preliminary experiments show that temporal information improves authentication accuracy. These empirical results are promising and justify further investigation.

1. Introduction

Biometric user authentication is typically formulated as a “one-shot” process, providing verification of the user when a resource is requested (e.g., logging in to a computer system or accessing an ATM machine). Once the user’s identity has been verified, the system resources are available for a fixed period of time or, more typically, until the user logs out or exits the session. While perhaps appropriate for short sessions or low-security environments, this model for authentication is flawed, as it is based on two strong assumptions: (1) a single verification is sufficient, and (2) the identity of the user is constant during the complete session. If the user leaves the work area for a while, or is forcibly removed in a hostile environment, the system continues to provide access to the resources that should be protected. Continuous biometrics attempts to improve on this situation by addressing these assumptions and making user authentication an ongoing process, rather than a one-time, point-of-access occurrence.

One way to approximate continuous biometrics is to require active user authentication on a regular basis, e.g., requesting a password or thumbprint verification every few minutes or so. In most environments, this is not an acceptable requirement. Passive verification, via modalities such as face recognition, can be used to authenticate at a much higher rate, perhaps several times per second, without requiring active user participation. This raises other questions that affect usability: What if, due to a lighting change, noise, or any of several other conditions, the verification fails momentarily? What if the modality in use cannot provide any authentication report for a time?

To be truly useful, continuous biometrics requires temporal integration. In general, a continuous biometric authentication system should be able to provide a meaningful estimate of authentication certainty at any time. This requires analyzing the temporal characteristics of biometric modalities and user behavior to provide a model of user identity that is a continuous function of time (or a discrete function with a reasonably small update rate). Intuitively, the certainty of an authentication result should be relatively high at the moment the score is reported (depending on the characteristics of the modality), and then decrease monotonically over time, until a new report is received.

Temporal integration is particularly relevant and useful in the case of multimodal biometrics. When multiple modalities are used in concert to provide user authentication, there is usually an implicit temporal model — even though the different modalities may report at slightly different times, the results are treated as if they had arrived simultaneously. This is equivalent to assuming a constant user model during this short period.

The most interesting and potentially useful case is when there are multiple modalities in use, where the characteristics of the various modalities may differ significantly.

For example, consider a high-security workstation situation where the biometric modalities are fingerprint, face, voice, and keyboard (keystroke pattern), representing a range of temporal characteristics (frequency and regularity of reports) and accuracies. Keystroke pattern recognition is likely to be the least reliable as an authentication technique, but at times it will give almost continuous output, while the other modalities may have nothing to report. Fingerprint recognition may be quite accurate, but will only be available occasionally. In this situation, we envision a system that monitors all the modalities and makes the best possible decision at any given point in time — even if there has been no information in the recent past. With this model of continuous authentication, a system can constantly communicate the degree of belief in a user's identity, and a monitoring system can implement an appropriate program of action for the particular security environment. A slight decline in authentication certainty may cause certain sensitive areas to be made inaccessible to the user (in many cases not at all disturbing the benign activity of the user), while a large decline may result in the system shutting down access.

Integrating biometric modalities into decision-making has produced successful results in terms of accuracy and robustness [1, 5, 8]. Still, this model of authentication fails to address the temporal nature of the problem. The main goal of this work is to present a temporal integration method to investigate potential benefits of time information for the realization of a continuous authentication system. As such, the system could generate continuous results in terms of confidence in the identity of the user, which would enable adjusting the security level accordingly in real time. In relation with behavioral traits, which are under investigation as admissible biometrics [7], temporal integration would be useful for detecting gradual or abrupt changes or variations in fitness to perform a task.

2. Multimodal Biometrics

There has been a good deal of research in recent years on integrating multiple modalities to identify or authenticate a user. In such a multimodal biometric system, the method of integration is very important, as the accuracy of a strong biometric could suffer when integrated with a weaker biometric [3, 6]. To our knowledge, there has been no published research in the biometrics community to date that focuses on temporal integration as formulated here.

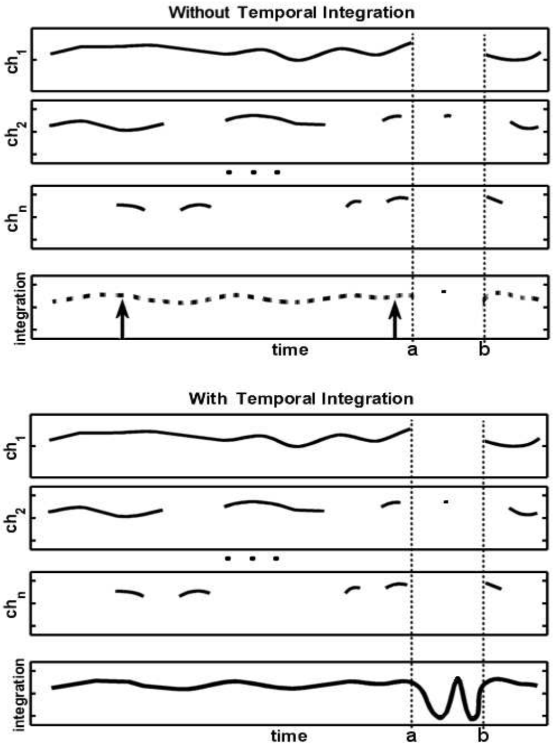


Figure 1: A static multimodal system (*top*) vs. one with temporal integration (*bottom*). Normalized scores from three channels are shown, with the integrated authentication score below. The multimodal system at top can not integrate information from all channels. For most of the time from a to b, the static multimodal system cannot perform authentication.

Figure 1 shows a qualitative comparison between a multimodal system that performs integration across modalities (without integration over time) and one which does temporal integration as well. The first system would be ineffective when there is no channel reporting — e.g., for most of the time between a and b. Through the entire sequence, the system would have to make decisions based on only partial observations, except where all channels are reporting an opinion (as indicated by arrows in Figure 1). In reality, due to the nature of biometric modalities involving lengthy computations or sample collection times, this should not be expected to happen frequently.

Interestingly, most accurate biometrics (iris scan, fingerprint, DNA matching and the like) are either lengthy procedures in collection or verification, or they are intrusive and cannot be performed frequently. A static multimodal system can only use such accurate indicators once they are observed.

2.1. Channel Integration

A multimodal biometric system can integrate modality information ("vertical" integration) at *feature*, *score*, and *figurethree* levels [1, 11, 5, 9]. In general, the most information is available at the feature level; integrating at this level is considered to be "early" integration. However, training at this level can be very complex and require an inordinate amount of data; later (higher) levels of integration are easier to build and often yield higher degrees of robustness. For decision level integration, it can be shown analytically that a strong biometric can achieve better accuracy alone than combined with a weaker biometric if both are operating at their cross-over points [6]. Unless the cross-over point of the weaker biometric is shifted, integration at the decision level would not be more accurate. Incorporating temporal information could change this limitation by shifting the cross-over point of weaker biometrics.

Since modality integration can be handled independent of temporal integration, it is possible to use various channel integration methods to improve overall accuracy of the system. In this work, channel integration is not our primary goal, so we chose a simple naive Bayes classifier to handle channel integration as a binary classification problem incorporating uncertainty measures. Similarity scores from individual biometric channels are normalized to the interval $[0, 1] \in \Re$ and integrated using the Bayes classifier. Our temporal integration method generates an expected score distribution and an estimated related uncertainty about this distribution. We weight class priors by the associated uncertainty before classification. It should be noted that weighting class priors would not scale well with larger data sets [4] presenting a potential limitation, especially since we are concerned with real-time operation.

2.2. Temporal Integration

There are several challenges for temporal ("horizontal") integration of a multimodal authentication system. First, as mentioned in the introduction, individual biometric channels cannot always provide simultaneous observations. One channel might provide information at a much higher frequency than another channel. Second, some channels might only provide sporadic observations over time. For example, we could not expect the user to provide a fingerprint at certain times. Third, for sporadic channels alone, temporal integration could be useless or statistically meaningless, if not impossible, to formulate, since there might be unexpectedly long intervals between observations. Fourth, the system should provide a way of making decisions during time intervals even if none of the individual channels provide any observations in that instant. For example, if we made observations δ milliseconds ago, then the system should be able to make decisions based on recent observations as we would not expect the user to be away in such a short interval. Our method addresses all of these challenges.

Logically, we have the choice of first integrating temporally or over channels (horizontally or vertically). If we first integrate over channels, then the problem is equivalent to temporal integration using a single biometric channel. On the other hand, integrating temporally first enables us to work with asynchronous biometric channels, since within some neighborhood in time of an observation we will have very good estimates from that observation. For making decisions in the absence of observations at a given point in time, we use expected values of observations from channels with varying degree of uncertainty. Perhaps the best approach, but also the most complex to formulate, is to integrate in both directions (across channels and across time) simultaneously, rather than sequentially.

3. Method

Just as in integrating channels, for temporal integration we can choose to integrate information at level of features, scores, or decisions. Our method works in continuous time by computing expected values of scores as a function of time difference between the last observation and current time. The main idea is based on the assumption that an authentication score is still valid for some amount of time, δt. As time passes, we should be less and less certain about this value. To formulate this idea as a function of time we estimate an uncertainty measure of scores per channel from the recent past, until a new observation is recorded. The joint posterior distribution of a score is approximated and then propagated over time until we obtain a new score from that channel. Due to the propagation of the score distribution over time, we use a degeneracy model for the uncertainty measure of each score.

The most important reason in favor of working with scores, rather than at the feature or decision level, is the way of modeling uncertainty of channel opinions. In lower levels, uncertainty has a related physical meaning. For example, at the physical measurement level, uncertainty is related to signal noise, which might not necessarily map well into an uncertainty about the decision. Treating scores as random variables is in fact this mapping, statistically backed by the Central Limit Theorem. Another reason to work with scores, aside from the underlying mathematical difficulty of using many features, is the fact that feature selection is still as much art as it is science. Naturally, we would prefer our integration method to be as general as possible. On the other hand, the later the integration, the more information is discarded, so early integration may achieve better results, using an appropriate set of features. After establishing promising results with scores, we plan to continue investigating such directions in the future.

Each channel is assumed to provide a normalized similarity score s, and an expected variance σ_{ch} as a characteristic parameter of the channel. If σ_{ch} is not provided, it is computed for each channel offline. This measure is equivalent to inherent uncertainty in a channel's decisions. This variance is only used as the default variance of the channel if computing the channel variance is not possible from recent past. For example, σ_{ch} is needed for initial few scores or for channels which provide scores at longer intervals. One might ask that if the uncertainty is known, why compute it from the past again? The reason is that the σ_{ch} measure itself varies over time. For example, if lighting conditions were the underlying reason for the face recognition channel to report highly variable scores over the past 5 seconds, this variability should be corrected in par with the lighting conditions.

We normalize channel scores to $[0, 1] \in \Re$, where 1 indicates perfect similarity to the user model and 0 indicates an unknown person. For channels with higher frequency, we compute the uncertainty σ_p from past scores within a τ_{ch} time period. Note that this duration is the crucial part of our method and it has a different value for each channel.

We model each channel with a Gaussian $\tilde{N}(\mu, \sigma_{ch})$ or $\tilde{N}(\mu, \sigma_p)$, where μ is the reported score for the channel, as discussed above. (We will refer to σ_{ch} and σ_p as σ from now on.) Consequently, scores are random variables with $s \sim \tilde{N}(\mu, \sigma)$. This distribution is propagated over time with increasing uncertainty in the score value as a function of time.

Figure 2 shows conceptually how a score s is treated. The darker lines over the Gaussian show the change in shape of Gaussian over time.

When a score is recorded, a timestamp t is generated and the uncertainty σ is computed over the past $t - \tau$, if applicable, otherwise $\sigma = \sigma_{ch}$. The idea is that we will be less and less certain about this score and probabilities of all possible scores will increase as time passes by.

The increase of uncertainty over time is computed as a function of time from the last score. We used an exponential degeneracy function $\phi(\tau)$ to estimate the mode ($\frac{1}{\sigma\sqrt{2\pi}}$) of the $\tilde{N}(\mu, \sigma)$ at $t + \tau$. The degeneracy function $\phi(\tau) = k \exp^{\alpha\tau}$ depends only on α which we take as the mean variability over the last τ_{ch} time period.

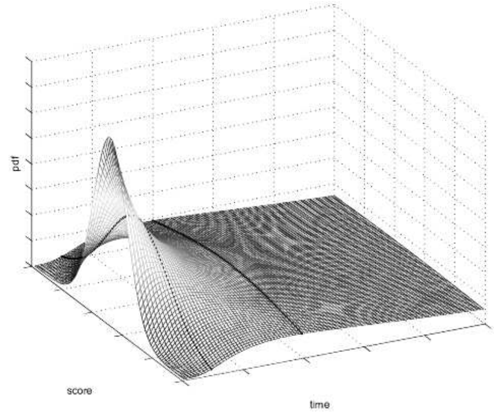


Figure 2: Propagation of scores and associated uncertainties over time. As time passes, σ increases from a recently computed σ_p.

Once an estimate of score distribution $\tilde{N}(\mu, \sigma)$ at $t+\tau$ is obtained, we compute the expected value of a score at $t + \tau$ from this distribution by evaluating

$$E_{\tilde{N}_{past}}\{N_{now}(s)\} = \int_{-\infty}^{\infty} \tilde{N}_{now}(s)\tilde{N}_{past}(s)ds$$

Note that the limits of the integral we are interested in are not $-\infty$ and ∞, but 0 and 1. Hence the distribution at $t + \tau$ is not a proper Gaussian anymore. However, the error resulting from ignoring the tails of this distribution is insignificant. Although we could opt for a proper distribution, such as a triangular distribution, this would introduce a larger modeling error. Alternatively, this Gaussian can easily be scaled to cover unit area, which would not change the expected value of the score. To evaluate the expected value we use the following approximation.

Suppose $X = \{X_1, X_2, ..., X_n\}$ is the set of random variables that characterize the model, with values $x_1, x_2, ..., x_n$. The expectation, $E(a)$, of a function $a(X_1, X_2, ..., X_n)$ can be approximated by

$$\sum_{x_1} ... \sum_{x_n} a(x_1, ..., x_n) P(X_1 = x_1, ..., X_n = x_n)$$

$$\approx \frac{1}{N} \sum_{k=0}^{N-1} a(x_1^k, ..., x_n^k)$$

where x_i^k are the values for point k in a sample of size N.

It should be noted that we want to minimize the filtering effect of our method, where occasional false positives and false negatives are *corrected* by subsequent scores. Therefore a predictor-corrector style modeling, such as a Kalman Filter, is not a model of choice. Also, the choice of the exponential function was based on lifetime modeling studies, which could be better modeled with $(1 - \tanh(x))$ or a similar function. The crucial heuristic of our method is the length of considered past, and how many correct scores it includes. Clearly, the degeneracy model leaves room for refinement. Incorporating contextual information successfully into the model and learning appropriate parameters from data are possible refinements.

4. Experiments

We chose face, voice, and fingerprint as individual biometric modes for simulating channels with different temporal characteristics. The lack of a suitable multimodal corpus with face recognition, voice verification, and fingerprints of individuals forced us to simulate individuals by matching independently collected data into virtual identities for 24 individuals. Scores from each channel are obtained as detailed below. Our goal is to achieve continuous multimodal authentication which is more accurate than the component channels and gives meaningful results at any point in time. A second set of experiments was run with different lengths of past scores in consideration.

4.1. Face Recognition

This is the channel with the highest reporting frequency. Face scores are obtained from a face recognizer based on Eigenfaces [12]. Images are obtained using a face detector built on [13] from 20fps video. For each individual, there is a 2 min video containing ~ 80 frames at (near) frontal pose. 20 images from frontal images were used for training. The data does not have frontal pose throughout the entire video sequence, hence the recognition does not provide good scores every $50ms$.

4.2. Voice Verification

A subset of the TIMIT database [10] was used. The subset contains LPC cepstrum feature vectors. The energy in all recordings was normalized to compensate for possible differences in loudness. After preemphasis, $16th$-order LPC-cepstra were calculated for $32ms$ frames centered at $16ms$ intervals. The feature vectors are the rows of the resultant matrix. Each frame is used as an independent sample drawn from the distribution of that speaker. Each speaker is modeled as a Gaussian. In total just under $15s$ of training data per speaker are available. Log-likelihoods are the scores for voice verification.

Table 1: Recognition rates of individual channels vs temporal multimodal integration.

Integrated	304	47.50%
Face	210	32.81%
Integrated	173	97.74%
Voice	171	96.61%

Table 2: Correct recognition at variable history lengths.

History length (secs)	0.5	1.0	2.0	5.0
Correct recognition	304	310	318	301
Recognition rate (%)	47.5	48.4	49.7	47.0

4.3. Fingerprint

A subset of fingerprint data was obtained from the FVC2002 fingerprint verification competition. A demo version of fingerprint identification/verification software [14] was used to obtain similarity scores between fingerprints. The software extracts minutiae-based features. It handles rotation and intensity variations. For successful operation it requires a minimum of 10 features for each fingerprint.

4.4. Results

We expect that temporal integration would be useful by enabling continuous authentication and by improving accuracy of a multimodal biometric system. Figure 3 shows decisions made by our method over a period of 32 seconds (each tick = 1 frame). The simulated user is the authentic (virtual) identity over the entire period, so that a 1 indicates a correct authentication, and a 0 marks where the system fails to authenticate the identity correctly. The varying face recognition scores are due to face motion, where it becomes frontal 6 times during the 32 second period. Better recognition scores are obtained when the face became full frontal in view.

The top three graphs show individual channel scores. The bottom graph shows the decisions obtained by our method with a history length of 0.5 second for all channels. The first few points are not affected by temporal integration due to insufficient history. In the case of a non-temporal multimodal system, all (if any) decisions would have to be based on what is observed at that point in time, regardless of what happened in the instant before. We can poll our system at any time for an authentication.

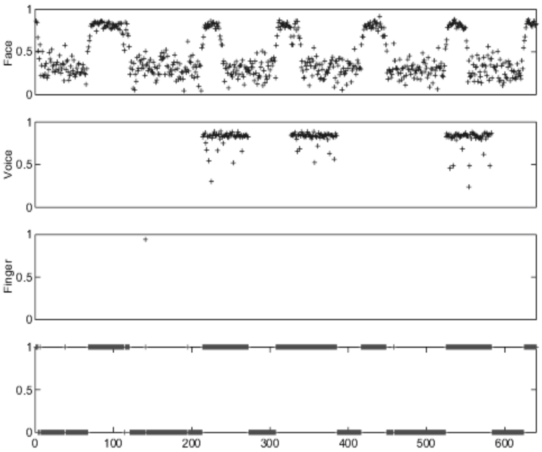


Figure 3: Temporal integration over a period of 32 seconds. Individual channels report scores in real time as they become available; note the single fingerprint score in frame 141. The bottom graph shows binary verification decisions made at every frame, a 1 **being valid authentication.**

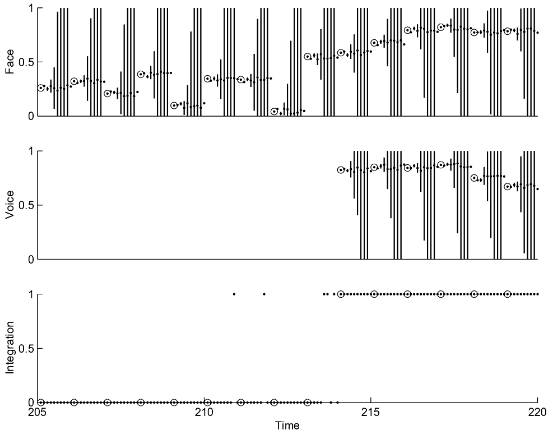


Figure 4: An enlarged version of Figure 3 between frames 205 and 220. Each frame is polled 10 times within the frame. Vertical lines show the variances of propagated distributions around the means since the last score. Fingerprint channel is not shown since the only score is beyond the history window of 0.5 seconds. Circles show actual scores from Figure 3.

To verify that integrated results are actually comparable to individual channel rates, we compared the correct recognition counts of integrated and individual channels. Table 1 shows this comparison over the periods when each individual channel is active.

Table 2 shows the effect of history length on recognition. The history length is applied to all channels. Our results suggest that there is a cross-over point for the length of relevant history, although more extensive study is necessary.

Figure 4 shows an enlarged sequence between frames 205 and 220 (0.75 seconds). Vertical lines show the variances of propagated distributions around the means since the last score. Fingerprint channel is omitted from both Figure 4 and Figure 5 since the only score lies beyond the relevant history of 0.5 seconds. An authentication result was requested 10 times within each frame. Our method is only limited by the underlying hardware in terms of temporal resolution, and an authentication score can be obtained given any point in time.

Figure 5 shows the same enlargement for a system that only integrates channels. Authentication is only possible when at least of the channels report an opinion. Note that in Figure 4 and Figure 5 the authentication is based only on face recognition scores for the first half of the sequence as no previous data was recorded from other channels within the last 0.5 seconds. Depending on the length of relevant history, our system can evaluate what has been seen within the last n seconds even if there were no scores reported from any channel, which would be impossible without temporal integration.

5. Conclusion

We have introduced a new model for temporal integration in biometric user authentication and developed an initial method for a continuous authentication system. Our temporal integration method depends on the availability of past observations, which makes the length of relevant history an important heuristic. Another important design choice is the degeneracy function. The existence of a cross-over point in the history suggests further investigation of the degeneracy.

We have shown on simulated data that our preliminary system can provide continuous authentication results which are consistently better than individual components of the system. Clearly, gathering a true multimodal database is very important for continued work in this field.

When the history length is set to 0, the system ignores temporal integration and degenerates into a multimodal system. Although our approach attempts to minimize the filtering effect of false positives and false negatives, our temporal integration method would suffer from this smoothing behavior to some degree as it stands. The net effect of this behavior is integration of positive decisions, as well as negative ones, as expected.

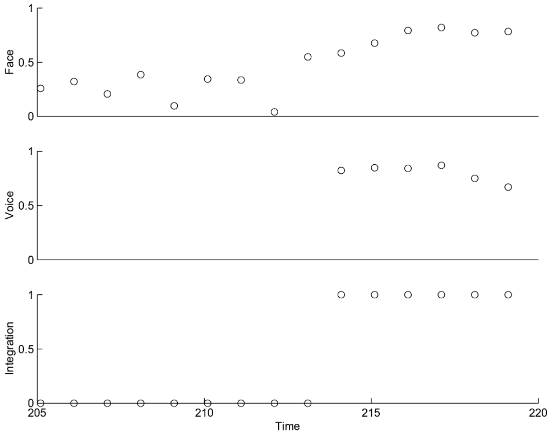


Figure 5: An enlarged version of Figure 3 between frames 205 and 220. Channel integration only, no temporal integration was performed. The system can perform authentication only when a score was reported by at least one channel.

References

[1] R. Brunelli and D. Falavigna, Person Identification using Multiple Cues, *IEEE Transactions on PAMI, Vol 12*, pp. 955-966, Oct. 1995.

[2] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, Multimodal Person Recognition using Unconstrained Audio and Video, *Second International Conference on AVBPA*, pp. 176-181, Washington D. C., USA, Mar. 1999.

[3] John Daugman, Biometric Decision Landscapes, *Technical Report, University of Cambridge, UK*, 1999.

[4] R. Duda, P. Hart, and D. Stork, Pattern Classification, *John Wiley & Sons, NY. 2nd Ed.*, 2001.

[5] L. Hong and A. Jain, Integrating Faces and Fingerprints for Personal Identification, *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 12*, pp. 1295-1307, Dec. 1998.

[6] L. Hong, A. K. Jain, and S. Pankanti, Can Multibiometrics Improve Performance?, *In Proceedings AutoID'99, NJ, USA*, pp. 59-64, Oct. 1999.

[7] A. K. Jain, R. Bolle, and S. Pankanti, Multimodal Biometrics: Personal Identification in a Networked Society, *Kluwer Academic Publishers*, pp. 1-38, 1999.

[8] N. Poh and J. Korczak, Hybrid Biometric Authentication System Using Face and Voice Features, *Third International Conference on AVBPA*, pp. 348-353, 2001.

[9] N. Poh, S. Bengio, and J. Korczak, A Multi-sample Multi-source Model for Biometric Authentication, *Proceedings, IEEE 12th Workshop on Neural Networks for Signal Processing*, pp. 375384, 2002.

[10] S. Seneff and V. Zue, Transcription and Alignment of the TIMIT Database, *In Proceedings of the Second Symposium on Advanced Man-Machine Interface through Spoken Language, Oahu, Hawaii*, Nov, 1988.

[11] G. Shakhnarovich, L. Lee, and T. Darrell, Integrated Face and Gait Recognition From Multiple Views, *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition, Lihue, HI*, Dec. 01.

[12] M. Turk and A. Pentland, Eigenfaces for Recognition, *Journal of Cognitive Neuroscience, Vol. 3, No. 1*, pp. 71-86, 1991.

[13] P. A. Viola and M. J. Jones, Robust Real-Time Object Detection, *Technical Report, COMPAQ Cambridge Research Laboratory, Cambridge, MA*, Feb. 2001.

[14] http://www.neurotechnologija.com/verifinger.html.

Color Correction for Face Detection Based on Human Visual Perception Metaphor

Krzysztof M. Kryszczuk and Andrzej Drygajło
Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne (EPFL)
[krzysztof.kryszczuk, andrzej.drygajlo]@epfl.ch

Abstract

In this paper we present a method of automatic color correction of face images and its application in a face detection algorithm. The color correction method is based on the phenomenon of color constancy observed in human visual perception. This technique is further applied in a face detection system, which draws upon the analogy to the parallel organization of visual neural pathways, the magno- and parvocellular channels. Presented method proved to be efficient in diverse background and illumination conditions, including face images with background chromatically close to human skin and where prominent facial features are obscured by adverse illumination conditions.

1. Introduction

Processing of human face images is an important research area with many applications, ranging from image enhancement to automatic face recognition in security systems. Beside the face itself, most face images contain background that must be discarded before subsequent face recognition process. Thus in most cases the first step in the image-processing task is the detection and localization of the face in the image.

A comprehensive overview of state-of-the-art face detection methods is presented by Yang et al. [1]. Particularly the knowledge-based, feature invariant, and template matching algorithms are listed as the most frequently used ones.

Human skin color can be regarded as an invariant feature and so are the skin color based methods classified by Yang et al. In fact, the skin color is an easily accessible, computationally inexpensive feature. Therefore it has been used in various face detection and recognition systems [2,3,4].

Despite the apparent skin color variations between different ethnic groups the actual skin chromaticity parameters can be clustered into a surprisingly compact set, which allows very accurate modeling [5]. The resulting skin color model can be used for color-based image segmentation focused on locating the skin-colored areas. This method of segmentation can deliver very precise distinction between the face and non-face areas of the image, provided that the background differs chromatically from the skin tone. The skin colored areas considered for further face recognition (verification or identification) can be accurately cropped out from the original image.

Skin model-based segmentation can result in precise skin area detection only if the model was created using the same spectral content of the skin illuminant as in the processed face image. Usually the information about skin illuminant is unknown for an arbitrary color image. Therefore a mismatch can occur between the model assumptions and the chromatic properties of skin depicted in the actual image. To avoid this mismatch it is necessary to normalize the image chromatically by introducing a chromatic frame of reference, common to both the model and the segmented image.

Precise retrieval of the spectral content of the illuminant in an arbitrary visual scene is an ill-posed problem [6]. Therefore a few heuristic methods have been proposed to normalize the chromaticity of the image [7].

Humans are known to cope well with the problem of color discrimination under varying illuminants thanks to the mechanism of color constancy observed in the natural visual processing [8]. In order to process a color face image acquired under unknown lighting conditions it is necessary to first employ a color correction mechanism, which would do what the phenomenon of color constancy does in humans.

The classical two assumptions that most color correction methods are based on are the “white world assumption” and the “gray world assumption” [9]. The first one assumes that there is a part of each image that is white. The second one postulates that all colors in the image should average to gray.

Hsu et al. [10] presented an interesting approach toward color correction of face images. They proposed an automatic color correction based on the localization of pixels with top 5% of luminance in the image, and assume those pixels to be 'white' (the "white world assumption"). Based on the chromatic distance between the white color and the actual color of the selected pixels the entire picture is being corrected. This method works with images that contain no specular reflections. However, in non-controlled environment or where the illumination control is limited, the specular reflections of the face appear very frequently.

The "gray world assumption" is not applicable to face images either, taking into consideration the fact that face images normally contain large skin-colored areas.

In this paper, we propose a new method of color cast removal from face images based on the inherent chromatic features of the face itself. In order to take full advantage of the method we incorporate it into a new robust face detection algorithm inspired by the organization of the human visual pathways (magno-and parvocellular channels) [8].

The rest of the paper is organized as follows: firstly, the general assumptions and details of the proposed method are explained. Then the proposed method is employed in a face detection algorithm. Results and final remarks conclude the paper.

2. The concept of image color correction inspired by the color constancy phenomenon

In order to be compliant with the assumption that the skin model must be built around a common frame of reference with processed face image we propose to use the chromatic information contained in the eye area as such a reference. We use this reference to perform the chromatic correction of the entire image. This process can be interpreted as a chromatic normalization.

The vast majority of images that are otherwise suitable for face verification (frontal pose, no occlusions etc.) show the face in such a way that both or at least one of the eyes are clearly visible. The image of an open eye contains normally the pupil, the iris, the eye-white and the eyebrow. A close inspection of eye images reveals that the eye-whites and the pupil areas are the locations, which are chromatically close to gray. The concept of the chromatic normalization can be best formulated as "bringing to gray what is closest to gray".

The proposed method is to find in the image of the eye pixels that are closest to gray. Consequently the chromatic coordinates of such pixels are modified to match gray, and same transformation is applied to the entire image. In order to perform this normalization procedure it is necessary to: localize the eye areas in the image, crop out the eye images and find the appropriate pixels for correction.

3. Color correction algorithm and creation of the skin color model

We build the skin color model using samples from face images from the VIDTIMIT database [11]. Before we take the samples, the images have to be chromatically normalized. To do that, we first locate the eye areas in the image. In our experiments we found them manually. We select for correction the area of left or right eye, whichever has the lower mean luminance. We assume that if the specular reflections are present, they will be more prominent in the overall "brighter" eye image.

For each pixel in the cropped eye image a distance from gray is calculated, using the formula:

$$D_g = \mathrm{abs}(R\text{-}G) + \mathrm{abs}(G\text{-}B) + \mathrm{abs}(\mathrm{B}\text{-}R), \qquad (1)$$

where D_g is the distance from gray and R,G,B are corresponding red, green and blue chromatic coordinates of the pixel. The pixel whose D_g is smallest is selected as the normalization reference and this pixel will be brought to gray. Next, the target gray coordinates C_g (equal for all three RGB channels) of the pixel are calculated as the rounded average of its actual coordinates:

$$C_g = \mathrm{round}(R+G+B) / 3. \qquad (2)$$

The difference between the original RGB coordinates of the pixel and its new target gray coordinates is calculated as follows:

$$\begin{aligned} D_R &= R - C_g, \qquad (3)\\ D_G &= G - C_g,\\ D_B &= B - C_g. \end{aligned}$$

The calculated values of D_R, D_G, *and* D_B are respectively subtracted from corresponding red, green and blue chromatic coordinates of every pixel in the original image. Should the resulting coordinate exceed the allowed range, its value is set to the extreme allowed value.

The described color correction was performed on 13 face images from the VIDTIMIT database. Then, from each image a 30 by 30 pixels patch containing skin from the face was cropped out. Each of the patches (initially in RGB format) has been converted into YCbCr color space, and the Y coordinate discarded. Resulting Cb and Cr coordinates have been clustered and their distribution modeled by a sum of two normal distributions (Figure 1).

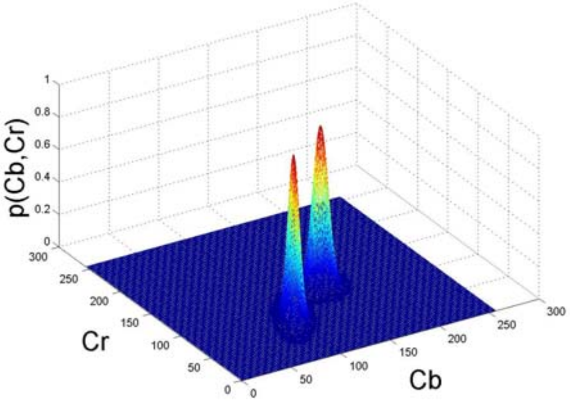


Figure 1. Skin color model in the YCbCr color space. The graph represents a probability density distribution of Cr and Cb coordinates of pixels that belong to skin-colored areas of the image.

4. Skin-color oriented image segmentation

For the processed image, the probability that each pixel's color belongs to the skin model distribution is calculated. The calculated probability values are stored in a new grayscale image, further referred to as the "skin map".

Performance of the model has been tested on a set of images different from those used for the creation of the skin color model. For each of the images the coordinates of the eyes were found manually, like during the model training. The test images were treated using the color correction procedure as described in Section 3. The model was tested for segmentation on images with and without the proposed color correction procedure. Example results are presented in Figure 2:

5. Application of the color correction method to face detection

Color information is used in many face detection and tracking algorithms. If all of the images originate from the same camera type and the spectral content of the illuminant is known, color-based segmentation is a way to quickly and robustly localize skin-colored areas without applying any prior chromatic correction. Typically, precise shape-based face detection techniques are applied after the color-based image segmentation [1].

However, if the face in the image is illuminated with a light source of unknown spectral power distribution, or/and the illumination is highly non-uniform, this approach often produces errors. Frequently the skin area in the image is not detected, or even worse, erroneously labeled.

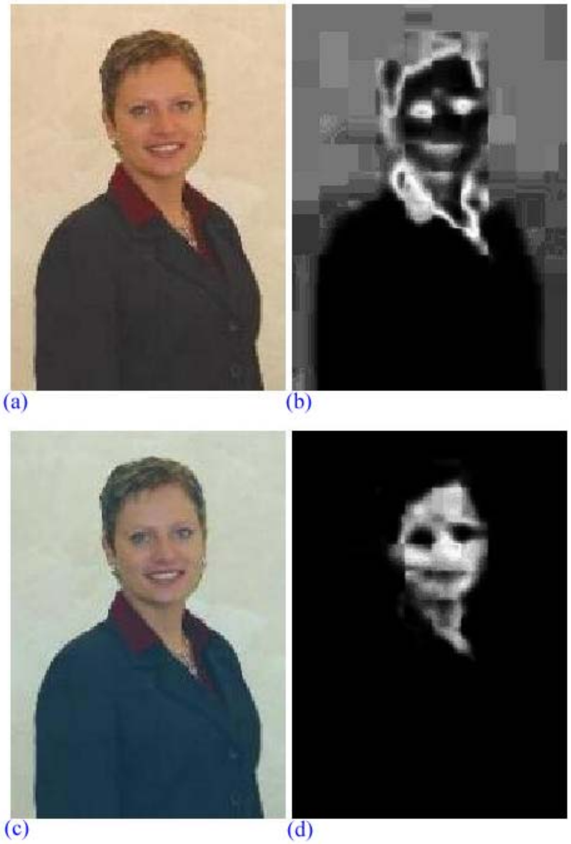


Figure 2. Results of the skin-color segmentation of the face images: (a) original image, (b) skin map of the original image, (c) original image after color correction, (d) skin map of the image after color correction.

In order to be able to use the color information to detect face in any image, we draw upon the analogy to the natural human visual system, which is known to successfully cope with the task of distinguishing colors in the presence of various illuminants.

Firstly, we revert to the idea of two separate neural pathways in the human visual system, the parvocellular and the magnocellular pathways [8] (further referred to as P-channel and M-channel, respectively). The M-channel conveys the generic shape, motion and intensity information, while the P-channel is responsible for the transmission of fine detail and color information.

As shown in numerous studies in visual search tasks, humans use the information from both neural pathways to find the desired information from a visual scene. For a given scene, the information from the channel that conveys the more discriminating data is used. If the object of interest stands out chromatically from the rest of the scene the color information is predominantly used. In a chromatically uniform scene the shape information prevails.

Therefore, we propose to use the color information simultaneously with shape-based face detection techniques for robust detection of faces in images as a high-level analogy to the M/P-channel visual processing in humans.

6. M/P-channel inspired face detection

Since the M- and P-channel processing is responsible for processing qualitatively different information about the image we propose to reproduce this dichotomy in a face detection system. In particular, we design a shape processing routine to model the M-channel, and a color processing routine to model the P-channel.

7.1 M-channel-based search

To model the M-channel search for faces in the visual scene (image) we use a template-matching approach. As a template a general grayscale 'average face' image is used (Figure 3).

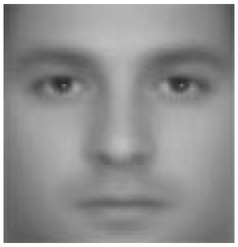


Figure 3. Average face template, resolution 115×119 (columns×rows).

The search process is performed as follows: the original image is converted into its grayscale version. Both the resulting grayscale image and the face template are high-pass filtered to reduce high contrasts in the face caused by non-uniform lighting distribution, specular reflections and self-shadows. Filtered image is divided into highly overlapping windows (5 pixels overlap) of the same size as the face template. For every window a 2D correlation coefficient with the face template is calculated. Negative correlation coefficient values are changed to null. Resulting values from the range (0,1) are regarded as probabilities of finding the face at a given window.

7.2 P-channel-based search

For each monochromatic window processed as described above, a corresponding window of identical size and location is cropped out of the original color image. Since each window is expected to contain a face image, we process them as if they would indeed contain a face. Figure 4 shows an example of this procedure. Figure 4(a) shows a chosen window before correction. Using the geometry of the average face template we automatically designate the areas that are most likely to contain the eyes in each window, presuming that the face is indeed there. Those areas are shown in Figure 4(c) and (d). Selection of the chromatic reference point for normalization is depicted in Figure 4(e). Consequently we perform color correction procedure described in Section 3, but the correction is applied to the current window only, rather than the entire image. The color-corrected window is shown in Figure 4(b).

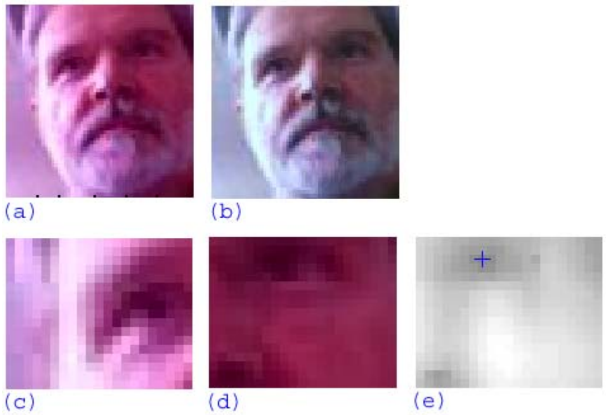


Figure 4. Automatic color correction of the image window; (a) original window, (b) window after color correction, (c) right eye area, (d) left eye area, (e) selection of the chromatic point of reference for color correction. Window taken from an image acquired from an USB camera (IBM), resolution 320×240. Window resolution 60×60.

Following the correction, a skin map is calculated for each window using the skin model as described in Sections 2 and 3.

Calculating the skin map for every window is a high computational burden. In order to speed up this stage, after the color correction step every window is downsampled by the factor of 4, and the skin map is calculated on the downsampled window.

The probabilities calculated for every pixel of the window are then averaged, which gives a mean likelihood measure that the given window contains the image of human skin.

7.3 Combining the M- and P-channel information

The procedure eventually returns two probability values for every window: P_L, probability that the shape in the window has a shape of a human face, and P_S, probability that the window contains object colored like human skin.

Since the information used in shape and color processing are obtained independently we calculate the joined probability that the window contains a face $P_{S,L}$ by multiplication of probabilities:

$$P_{S,L} = P_S \cdot P_L. \tag{4}$$

The window with the highest $P_{S,L}$ is a candidate to be the actual detected face in the image. However, the exact size of the face in the image is not *a priori* known, so it is necessary to perform the face search as described above for a few scaled versions of the face template. For each run with a different template size, we obtain a new $P_{S,L}$ and the window that corresponds to it. We choose the window wit the highest overall value of the probability $P_{S,L}$.

The presented method of color correction for face detection has been tested on high quality images from the VIDTIMIT database, pictures with adverse lighting conditions taken from a web-cam, and scanned photographs. Figures 5-9 show the results of the experiments. Figure 5 shows an example of a good quality picture taken from the VIDTIMIT database. Figures 6, 7 and 8 show the images acquired from a computer USB camera (IBM), taken in our laboratory, where the walls and the ceiling are chromatically close to the color of the skin. The face in Figure 6 is illuminated from its right side with daylight (coming from a window). Due to this condition the right side of the face shows strong reflections while the left side remains in the shadow. Figures 7 and 8 have the same daylight illuminant as present in Figure 6, additionally augmented by warm-white light originating from the fluorescent lamps overhead. In those pictures, top left part of the head shows highlights and the entire scene is illuminated by sources of two distinctly different spectral contents. Finally, Figure 9 shows a picture scanned from a paper photograph and saved in low resolution. In this figure, the background is chromatically very close to the skin tone.

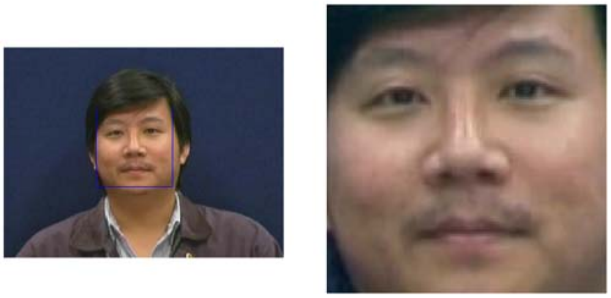


Figure 5. Image from VIDTIMIT database (res. 512×384)

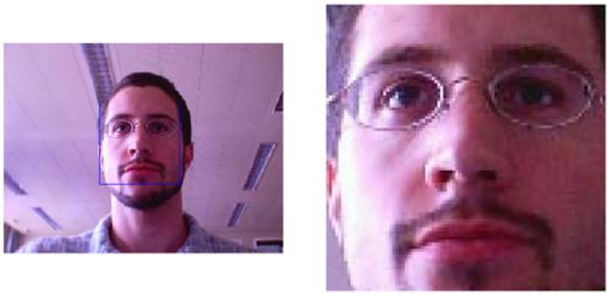


Figure 6. Image acquired from an USB camera (res. 320×240)

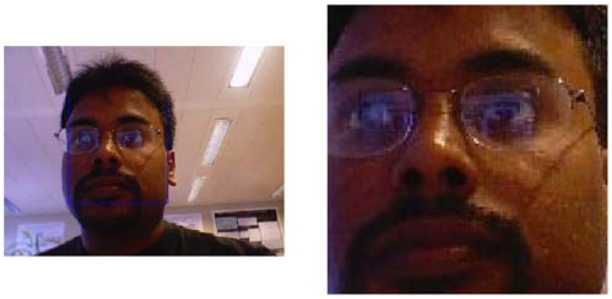


Figure 7. Image acquired from an USB camera (res. 320×240)

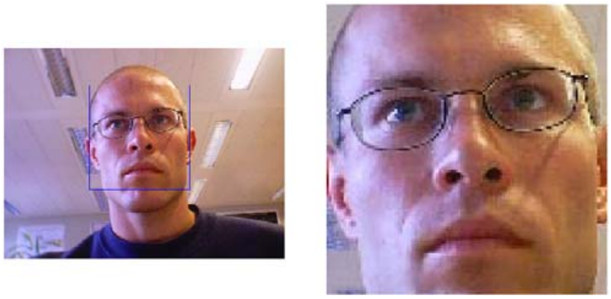


Figure 8. Image acquired from an USB camera (res. 320×240)

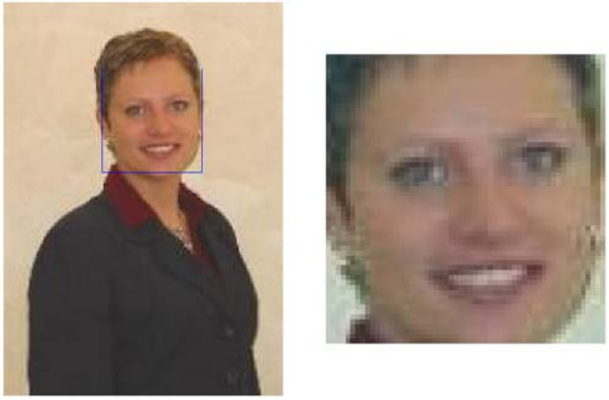


Figure 9. Image scanned from paper photograph, resolution of the jpeg compressed image 157×221.

7. Conclusions

In this paper we propose a method that successfully performs color correction of face images. We presented a way to incorporate this method into a generic algorithm that detects faces in images of various resolution and quality, where the face image may be distorted by adverse illumination. The advantage of the technique is that it detects a face if it is present; if it is not this fact can be inferred from the probability measures obtained during the detection process.

The algorithm may produce erroneous detection only in rare cases where neither the shape, nor the color can deliver reliable information about the location of the face. This can happen when the shape of the face is heavily distorted by adverse lighting conditions and at the same time the color of the background is indistinguishable from the skin tone. In such cases, due to lack of reliable color clues the system relies entirely on the template matching to find the best face candidate. In order to improve the system performance in such cases more appropriate filtering method than simple high-pass filter should be applied.

8. References

[1] M.H. Yang, D.J. Kriegman, N. Ahuja, “Detecting Faces in Images: A Survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, No. 1, January 2002.

[2] S. Marcel, S. Bengio, “Improving Face Verification using Skin Color Information”*, Proc. 16th International Conference on Pattern Recognition*, 2002.

[3] M. Störring, T. Kočka, H.J. Andersen, E. Granum, “Tracking regions of human skin through illumination changes”, *Pattern Recognition Letters,* No. 24, 2003, pp. 1715-1723.

[4] A.W. Senior, “Face and feature finding for a face recognition system”, *Proc. 2nd International Conference on Audio- and Video-based Biometric Person Authentication*, Washington D.C, 1999, pp. 154-159.

[5] E. Angelopoulou, R. Molana, K. Daniilidis, “Multispectral Skin Color Modeling”, Technical Report MS-CIS-01-22, June 22, 2001.

[6] R. Gross, V. Brajovic, “An Image Preprocessing Algorithm for Illumination Invariant Face Recognition”, *4th Intl Conf. On Audio- and Video-Based Biometric Person Authentication*, Guilford, UK, 2003, pp.10-17.

[7] M. Störring, H.J. Andersen, E. Granum, “Estimation of the Illuminant Colour from Human Skin Colour”, *4th Intl. Conf. On Automatic Face and Gesture Recognition*, Grenoble, France, 2000, pp. 64-69.

[8] M.S. Gazzaniga, *Cognitive Neuroscience, the Biology of the Mind*, W.W. Norton, New York 2002.

[9] A.C. Hurlbert. *The Computation of Color*. PhD thesis, Massachusetts Institute of Technology, Sept. 1989.

[10] R.L. Hsu, M. Abdel-Mottaleb, A.J. Jain, “Face Detection in Color Images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, May 2002, pp. 696-706.

[11] C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes", *Pattern Recognition Letters*, Vol. 24, No. 14, 2003, pp. 2409-2419.

Low-Dimensional Image Representation for Face Recognition

Jongmoo Choi and Juneho Yi
School of Information & Communication Engineering
Sungkyunkwan University
300 Chunchun-Dong, Jangan-Gu Suwon 440-746, Korea

Abstract

We present two novel methods for extremely low-dimensional representation of facial images that achieve graceful degradation of recognition performance. We have observed that if data is well-clustered into classes, features extracted from a topologically continuous transformation of the data are appropriate for recognition when low-dimensional features are to be used. Based on this idea, our methods are composed of two consecutive transformations of the input data. The first transformation is concerned with best separation of the input data into classes and the second focuses on the transformation that the distance relationship between data points before and after the transformation is kept as closely as possible. We employ LDA (Linear Discriminant Analysis) for the first transformation, and SOFM (Self-Organizing Feature Map) or MDS (Multi-Dimensional Scaling) for the second transformation. We have evaluated the recognition performance of our methods: LDA combined with SOFM method and LDA combined with MDS method. Experimental results using Yale, AT&T and FERET facial image databases show that the recognition performance of our methods degrades gracefully when low-dimensional features are used.

1. Introduction

In computer vision research, dimensional reduction is one of the most important problem. Especially, in face recognition research, statistical methods for feature extraction such as PCA (Principal Components Analysis) [1] [2], ICA (Independent Components Analysis) [3] [4] and LDA (Linear Discriminant Analysis) [5] [6] are widely used for dimensional reduction. The problem on extremely low-dimensional image representation for face recognition has little been investigated while many researchers study on face recognition robust to illumination [7] [8],posture [9] and facial expression changes [10]. When facial feature data need to be stored in low capacity storing devices such as bar codes and smart cards, extremely low-dimensional image representation of facial data is very important.

In this research, we present two novel methods for low-dimensional data representation of which the recognition performance degrades gracefully. The technique reduces dimension of high-dimensional input data as much as possible, while preserving the information necessary for the pattern classification. The algorithms like PCA, LDA and ICA can be used for reduction of the dimension of the input data but are not appropriate for low-dimensional representation of high dimensional data because their recognition performance degrade significantly. For low-dimensional data representation, SOFM (Self-Organizing Feature Map) [11], PP (Projection Pursuit) [12] and MDS (Multi-Dimensional Scaling) [13] are proposed. These techniques suitable for data representation in low-dimensions, usually two or three dimensions. They try to represent the data points in a such way that the distances between points in low-dimensional space correspond to the dissimilarities between points in the original high dimensional space. However, these techniques do not yield high recognition rates mainly because they do not consider class specific information. Our idea is that these methods incorporated with class specific information can provide high recognition rates.

We have found that if data is well-clustered into classes, features extracted from a topologically continuous transformation of the data are appropriate for recognition when extremely low-dimensional features are to be used. Based on this idea, we first apply a transformation to the input data to achieve the most separation of classes, and then apply another transformation to maintain the topological continuity of the data that the first transformation produces. By Topological continuity [11], we mean that the distribution of data before and after dimensional reduction is similar in the sense that the distance relationship between data points is maintained.

To experimentally prove our claim, we have proposed two novel methods for extremely low-dimensional representation of data with graceful degradation of recognition performance. It is composed of two consecutive transformations of the input data. The first transformation is concerned with best separation of the input data into classes and the second focuses on the transformation in the sense that the distance relationship between data points is kept. Our methods are implemented as the following. The first method employs LDA and SOFM for the transformations. SOFM preserves the distance relationship before and after the data is transformed. This way, it is possible to represent data in low-dimensions without serious degradation of recognition performance. The second method applies LDA and classical MDS. The MDS preserves the distance relationship before and after the data is transformed as closely as possible.

The following section gives a brief overview of the feature extraction and dimensional reduction methods that have preciously been used for object recognition. In section 3, we describe the proposed LDA combined with SOFM method and the LDA combined with MDS method, respectively. (Let us call them 'LDA+SOFM' and 'LDA+MDS' methods, respectively.) We report the experimental results on the recognition performance of LDA+SOFM and LDA+MDS methods in section 4.

2. Dimensional Reduction and Topological Continuity

Facial images of high resolution exhibit significant correlation between neighboring pixels. There have been reported many algorithms for dimensional reduction and feature extraction. Dimensional reduction methods can be categorized into *topologically continuous map* and *topologically discontinuous map* methods. Among the former methods are SOFM, MDS and GTM (Generative Topographic Mapping) [14] and these methods are used mainly for data visualization. LDA, Kernel LDA [15] and multi-layer neural networks are examples of the latter category and are mostly used for pattern classification [16].

2.1. Difficulty of Extremely Low-Dimensional Data Representation

We can achieve very low-dimensional data representation with graceful degradation of performance by using a *topologically continuous map* method when the data is well clustered into classes. However, the typical facial image data in real environments do not have well-clustered distribution as shown in Fig. 1. Fig. 1 shows an example that within-class variance is much higher than between-class variance. In such case, it is not guaranteed to achieve high classification performance by a *topologically continuous map* method although we can get a low-dimensional data set. Accordingly, we have to focus more on the discriminant power rather than dimensional reduction in the case of Fig. 1. Since LDA yields a linear transformation that minimizes within-class variations while maximizing between-class variations, we can apply LDA to facial images in real environments [5] [6].

In an LDA method, the dimension of feature space is related to the number of classes. It means that we might not be able to achieve dimensional reduction lower than the dimension of input space depending on the number of classes. In addition, blind dimensional reduction using just a few basis vectors that correspond to large eigenvalues drastically degrades the recognition rate [17].

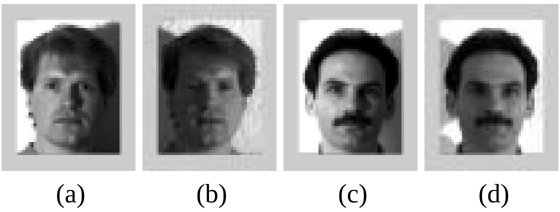


Figure 1: Facial images in the case of illumination changes. The images show that within-class variances are much higher than those between-class variances. For example, the cosine distance between (a) and (b) is 0.622 though they are from the same person. On the other hand, the cosine distance between (a) and (c) is 0.933 though they are from different persons. The value closer to 1.0 represents more similarity in the case of cosine distance.

3. Our Methods for Low-Dimensional Data Representation

3.1. Two-Stage Dimensional Reduction

We present two methods for extremely low-dimensional data representation by applying two different transformations in a row. The first stage is only concerned with best separation of classes. Once the data is rendered well-separated into classes by the first stage transformation, the second stage transformation only focuses on preservation of topological continuity before and after the transformation of the data. As previously described, the idea is based on the fact that if data is well-clustered into classes, features extracted from a topologically continuous transformation of the data are appropriate for recognition when extremely low-dimensional features are to be used. Fig. 2 illustrates the idea of our method. In the example, the two-stage dimensional reduction method solely makes a low-dimensional feature space appropriately for classification.

3.2. Method I: LDA+SOFM

Let us $\mathbf{x}_k \in \mathbb{R}^N, k = 1, \cdots, M$ be a set of training data. LDA produces a linear discriminant function $\mathbf{f}(\mathbf{x}) = \mathbf{W^T x}$ which maps the input data onto the classification space. We have employed FLD (Fisher's linear discriminant) as an instance of LDA techniques. FLD finds a matrix $\mathbf{W}$ that maximizes

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} \tag{1}$$

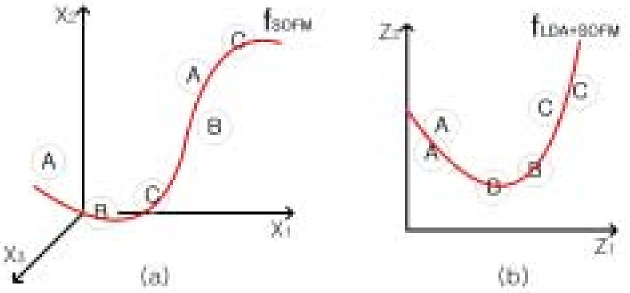


Figure 2: Conceptual illustration of dimensional reduction of 3D data into 1D data: (a) shows the input data distribution in a 3D input space. The curve represents interpolated weight vectors of a trained SOFM. Although we can reduce its dimension into 1D using the SOFM, the data in 1D feature space become not clustered. An LDA-like method can map 3D input space onto 2D space so that the data may be well classified as shown in the figure (b). If any basis vector of the 2D space were eliminated, we would not classify A, B and C appropriately using the 1D data projected onto the remaining axis.

where

$$\mathbf{S}_b := \frac{1}{M}\sum_{i=1}^{M}(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T \tag{2}$$

$$\mathbf{S}_w := \frac{1}{M}\sum_{i=1}^{C}\sum_{\mathbf{x}\in\chi_i}(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T. \tag{3}$$

$\mathbf{S}_b$ and $\mathbf{S}_w$ are between- and within-class scatter matrices, respectively. χ_i represents i^{th} class and the mean of class χ_i, $\mathbf{m}_i$ is computed as $\mathbf{m}_i := \frac{1}{n_i}\sum_{\mathbf{x}\in\chi_i}\mathbf{x}$. $\mathbf{m}$ denotes the total mean. $\mathbf{W}$ is computed by maximizing $J(\mathbf{W})$. That is, we find a subspace where, for the data projected onto the subspace, between-class variance is maximized while minimizing within-class variance. As a result of the first transformation, we obtain $\mathbf{z} = \mathbf{W}^T\mathbf{x}$.

After the stage of LDA, the next stage maps $\mathbf{z}$ onto a low-dimensional feature space $\mathbf{f} = \mathbf{G}(\mathbf{z})$ by SOFM. SOFM is a kind of competitive network. SOFM first determines the winning neuron using a competitive layer. Next, weight vectors for all neurons within a certain neighborhood of the winning neuron are updated using the Kohonen rule [11]. When a vector is presented, the weights of the winning neuron and its neighbors move toward the input pattern. After learning, the neurons of the output layer will be a feature map revealing a distance relationship within input patterns.

3.3. Method II: LDA+MDS

3.3.1 Classical MDS

Given M points and the corresponding dissimilarity matrix, classical MDS is an algebric method to find a set of points in low-dimensional space so that the dissimilarity are well-approximated by the interpoint distances. Let us $\mathbf{x}_k \in \mathbb{R}^N, k = 1, \cdots, M$ be a set of observations and

$$\mathbf{D} = \begin{pmatrix} d_{11} & \cdots & d_{M1} \\ \vdots & \ddots & \vdots \\ d_{1M} & \cdots & d_{MM} \end{pmatrix} \tag{4}$$

be a dissimilarity matrix, where d_{ij} is a squared Euclidean distance, $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{y}_j\rangle$. In summary, the inner product matrix of raw data $\mathbf{B} = \mathbf{X^TX}$ can be computed by $\mathbf{B} = -\frac{1}{2}\mathbf{HDH}$, where $\mathbf{X}$ is the data matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_M] \in \mathbb{R}^{N\times M}$ and $\mathbf{H}$ is a centering matrix $\mathbf{H} = \mathbf{I} - \frac{\mathbf{1}}{\mathbf{M}}\mathbf{1^T1}$. $\mathbf{B}$ is real, symmetric and positive semi-definite. Let the eigendecomposition of $\mathbf{B}$ be $\mathbf{B} = \mathbf{V\Lambda V^T}$, where $\mathbf{\Lambda}$ is a diagonal matrix and $\mathbf{V}$ is a matrix whose columns are the eigenvectors of $\mathbf{B}$. The matrix $\hat{\mathbf{X}}$ for low-dimensional feature vectors can be obtained as

$$\hat{\mathbf{X}} = \mathbf{\Lambda_k^{1/2}V_k^T} \tag{5}$$

where $\mathbf{\Lambda}_k^{1/2}$ is a diagonal matrix of k largest eigenvalues and $\mathbf{V}_k$ is its corresponding eigenvectors matrix. Thus, we can compute a set of feature vectors, $\hat{\mathbf{X}}$, for a low-dimensional representation. See [18] for a detailed description.

3.3.2 Mapping onto an MDS subspace via PCA

We could not map new input vectors to features by using the classical MDS because the map is not explicitly defined in the classical MDS [19]. We used a method that achieves mapping onto an MDS subspace via PCA based on the relationship between MDS and PCA. Let $\mathbf{Y_{MDS}}$ be a set of feature vectors in an MDS subspace and $\mathbf{Y_{PCA}}$ be a set of feature vectors in a PCA subspace. Let $\mathbf{\Lambda_{MDS}}$ denotes the digonal matrix of eigenvalues of inner product matrix $\mathbf{B}$. Then, the relationship between PCA and MDS is

$$\mathbf{Y_{PCA}} = \mathbf{\Lambda_{MDS}}^{1/2}\mathbf{Y_{MDS}}. \tag{6}$$

The derivation of equation (6) is described in the following [20]. For centered data, the covariance matrix is $\mathbf{\Sigma} = \mathbf{E}\{\mathbf{XX^T}\} = \frac{1}{M}\mathbf{XX^T}$. PCA is concerned with the eigendecomposition of the covariance matrix as follows;

$$\mathbf{\Sigma V_{PCA}} = \frac{1}{M}\mathbf{XX^TV_{PCA}} = \mathbf{V_{PCA}\Lambda_{PCA}}. \tag{7}$$

MDS is concerned with the eigendecomposition of the inner product matrix $\mathbf{B} = \mathbf{X^TX}$ as follows;

$$\mathbf{BV_{MDS}} = \mathbf{X^TXV_{MDS}} = \mathbf{V_{MDS}\Lambda_{MDS}}. \tag{8}$$

Using equations (7) and (8), we have

$$\mathbf{X}\mathbf{X}^{\mathbf{T}}(\mathbf{X}\mathbf{V}_{\mathbf{MDS}}) = (\mathbf{X}\mathbf{V}_{\mathbf{MDS}})\mathbf{\Lambda}_{\mathbf{MDS}} \tag{9}$$

and $\mathbf{V}_{\mathbf{PCA}} = \mathbf{X}\mathbf{V}_{\mathbf{MDS}}$, where $\mathbf{\Lambda}_{\mathbf{PCA}} \simeq \mathbf{\Lambda}_{\mathbf{MDS}}$. The feature vector set of PCA subspace is

$$\begin{aligned} \mathbf{Y}_{\mathbf{PCA}} &= \mathbf{V}_{\mathbf{PCA}}^{\mathbf{T}}\mathbf{X} \\ &= (\mathbf{X}\mathbf{V}_{\mathbf{MDS}})^{\mathbf{T}}\mathbf{X} \\ &= \mathbf{V}_{\mathbf{MDS}}^{\mathbf{T}}\mathbf{B} \\ &= \mathbf{\Lambda}_{\mathbf{MDS}}\mathbf{V}_{\mathbf{MDS}}^{\mathbf{T}} \\ &= \mathbf{\Lambda}_{\mathbf{MDS}}^{\frac{1}{2}}\mathbf{Y}_{\mathbf{MDS}}. \end{aligned} \tag{10}$$

Note that, whereas the classical MDS computes inner product matrix $\mathbf{B}$ from the given dissimilarity matrix $\mathbf{D}$ without using input patterns $\mathbf{X}$, in this dimensional reduction problem for pattern recognition, we can obtain $\mathbf{B}$ directly from the input patterns $\mathbf{X}$. For the purpose of low-dimensional feature extraction, we need to compute projections onto LDA and MDS subspaces. Let $\mathbf{p}$ be an input pattern, then the feature vector in LDA+MDS space becomes

$$\mathbf{f}_{\mathbf{LDA+MDS}} = (\mathbf{\Lambda}_{\mathbf{PCA}}^{-\mathbf{1/2}})\mathbf{W}_{\mathbf{PCA}}^{\mathbf{T}}\mathbf{W}_{\mathbf{FLD}}^{\mathbf{T}}\,\mathbf{p}. \tag{11}$$

4. Experimental Results

We have evaluated the recognition performance of the proposed LDA+SOFM and LDA+MDS methods as follows.

4.1. Experiment I: LDA+SOFM with Yale and AT&T Databases

We have compared the recognition performance of PCA [2], LDA [6], SOFM and the proposed LDA+SOFM method using three different facial image databases.

4.1.1 Facial Image Databases

We have used Yale [21] and AT&T [22] databases. The Yale database consists of facial images captured in simple backgrounds. Facial images are gathered under variations of luminance, facial expressions, glasses and time intervals. The database contains 165 images of 15 persons. The facial images of the AT&T database are gathered under variations of postures. The database contains 400 images of 40 persons. We tightly cropped and normalized all the facial images in each database for the experiment.

4.1.2 Training and Testing

In the FLD stage, we compute the linear transformation matrix for FLD and then transform the entire patterns in training sets into feature vectors using the matrix. In the test stage, each input pattern was also transformed into its corresponding feature vector using that matrix. We have used a nearest neighbor classifier for recognition.

Table 1: Correct Recognition Rates (%) (C: number of class)

		Yale	AT&T
Dimension	Methods	(C=15)	(C=40)
2	PCA	16.4	11.9
	LDA	41.8	11.9
	SOFM	64.3	71.3
	LDA+SOFM	96.4	86.2
C-1	PCA	87.3	94.0
	LDA	98.2	94.8

In the SOFM stage, the entire training patterns are represented by the indices of neurons corresponding to two-dimensional map. In testing, only the node that is the most similar to the given input pattern is activated. As a result, input patterns are classified into classes of the activated nodes. In the proposed method, the number of input neurons in SOFM is the same as the dimension of feature vectors obtained from the FLD stage. The output layer represents a two dimensional square map. Table 1 shows the recognition performance in the case of dimensional reduction to two dimensions.

4.1.3 Cross Validation

The performance of the SOFM algorithm varies depending on the initial parameters. Hence, we have applied cross validation to correctly evaluate the performance of SOFM. We have partitioned the training set into two subsets. One set is for learning and the other for validation. First, we change the number of grids. After learning using multiple SOFMs, we evaluate the performance using the validation set. We have decided the number of neurons as the number of grids that have the highest average recognition performance. Secondly, after the number of neurons is settled, multiple SOFMs with various initial parameters are learned by the learning set. Then we select the SOFM that has high performance corresponding to the upper 10% in the validation set.

4.1.4 Results

We show initial experimental results for extreme dimensional reduction to two dimensions. As shown in Table 1, LDA+SOFM method performs better than the others in the case of very low-dimensional representation. The recognition rate of LDA is high (98.2%) when a sufficient number, C-1, of features are used. However, the recognition rate degraded significantly to 41.8% when only two dimensional representation of the data is used. The recognition rate of SOFM is higher than that of LDA when two dimensional representation is employed. The experimental results show that very low-dimensional data representation with graceful degradation of recognition performance can be achieved by using a topologically continuous transformation after the input data is rendered well clustered into classes.

4.2. Experiment II: LDA+MDS with FERET Database

We have compared the recognition performance of LDA [6] and the proposed LDA+MDS method using a part of FERET database [23].

4.2.1 FERET Database and Experimental Method

The FERET Database is a set of facial images collected by NIST from 1993 to 1997. For preprocessing, we closely cropped all images in the database which include internal facial structures such as the eyebrow, eyes, nose, mouth and chin. The cropped images do not contain the facial contours. Each face image is downsampled to 50x50 to reduce the computational complexity and histogram equalization is applied.

The whole set of images, U, used in the experiment, consists of three subsets named 'ba', 'bj' and 'bk'. Basically, the whole set U contains images of 200 persons and each person in the U has three different images within the 'ba', 'bj' and 'bk' sets. The 'ba' set is a subset of 'fa' which has images with normal frontal facial expression. The 'bj' set is a subset of 'fb'. The images of 'fb' have some other frontal facial expressions. The 'ba' and 'bj' set contain 200 images of 200 persons, respectively. The 'bk' set is equal to the 'fc' of which images were taken with different cameras and under different lighting conditions. The 'bk' set contains 194 images of 194 persons.

For the experiment, we have divided the whole set U into training set (T), gallery set (G) and probe set (P). In order to get an unbiased result of performance evaluation, no one within the training set (T) is included in the gallery and the probe sets. i.e. $T \cap \{G \cup P\} = \emptyset$. The experiment consists of two sub-experiments; The first experiment is concerned with evaluation regarding normal facial expression changes. We use the 'ba' set as the gallery and the 'bj' set as the probe. The second experiment is to evaluate the performance under illumination changes. We have assigned the 'ba' set to the gallery and the 'bk' set to the probe. In addition, we randomly selected 50% of the whole set in each sub-experiment in order to reduce the influence of a particular training set because a facial recognition algorithm based on statistical learning depends on the selection of training images. Thus, a training set contains 100 persons in each sub-experiment.

We have compared the recognition performance of our LDA+MDS with that of LDA. In each algorithm, we have computed a linear transformation matrix that contains a set of basis vectors for a subspace using the training set, and then have transformed the entire patterns in the gallery set into feature vectors. For the test, each input pattern in the probe set was transformed into its corresponding feature vector. We used a nearest neighbor classifier for recognition.

4.2.2 Results

As shown in Figures 3 and 4, LDA+MDS method performs better than the others in the case of low-dimensional representation. The experimental results show that low-dimensional data representation with graceful degradation of recognition performance can be achieved by using an inter-distance preserving transformation after the input data is rendered well clustered into classes. The recognition rate for a given number of features in these figures was obtained by averaging thirty experiments.

Figures 5 and 6 show the recognition rates of LDA+MDS for three different distance measures, L1, L2 and cosine. We can see that there is no significant performance difference between the three distance measures.

5. Conclusion

This research features novel methods for low dimensional reduction of facial data that do not give significant degradation of the recognition rate. The LDA+SOFM method achieves very accurate recognition rates although only two dimensional features are used for recognition. The LDA+MDS method also outperforms LDA method when represented in a low-dimensional space. These results experimentally prove that if data is tightly clustered and well separated into classes, a few features extracted from a topological continuous mapping of the data are appropriate low dimensional features for recognition without significant degradation of recognition performance.

Our methods are practically useful for face recognition, especially when facial feature data need to be stored in low capacity storing devices such as bar codes and smart cards. It is also readily applicable to real-time face recognition in the case of a large database.

Acknowledgments

This work was supproted in part by grant No. R01-1999-000-00339-0 from the Basic Research Program KOSEF and BK21.

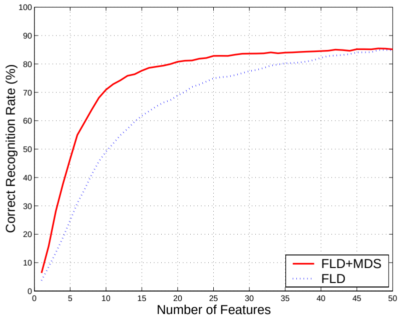


Figure 3: Comparison of recognition rates for ‘ba’-‘bj’ set

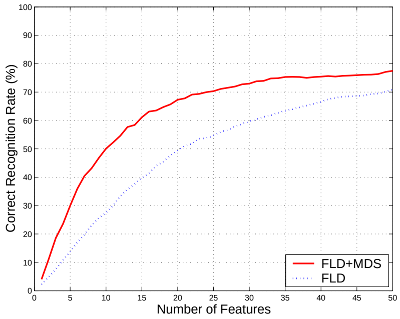


Figure 4: Comparison of recognition rates for ‘ba’-‘bk’ set

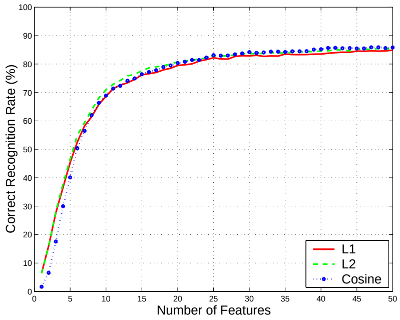


Figure 5: Comparison of recognition rates for various distance measures in the case of ‘ba’-‘bj’ set

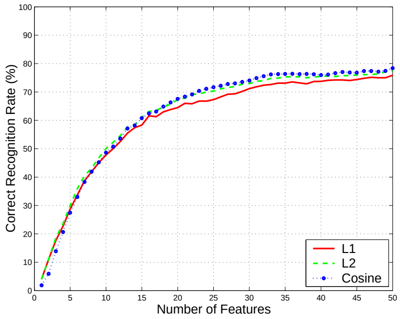


Figure 6: Comparison of recognition rates for various distance measures in the case of ‘ba’-‘bk’ set

References

[1] Kirby, M., Sirovich, L.: “Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces,” *IEEE Trans. on PAMI*, vol. 12, no. 1, pp. 103–108, 1990.

[2] Turk, M., Pentland, A.: “Eigenfaces for Recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] Bartlett, M.S., Martin, H., Sejnowski, T.J.: “Independent Component Representations for Face Recognition,” *Proceedings of the SPIE*, Vol. 3299, pp. 528–539, 1998.

[4] Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*, John Wiley & Sons, Inc., 2001.

[5] Etemad, K., Chellappa, R.: “Discriminant Analysis for Recognition of Human faces image,” *Journal of Optical Society of America*, vol. 14, no. 8, pp. 1724–1733, 1997.

[6] Belhumeur, P., Hespanha, J., Kriegman, D.: “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,” *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[7] Bischof, H., Wildenauer, H., Leonardis, A.: “Illumination Insensitive Eigenspaces,” *Eighth International Conference on Computer Vision*, Vol. 1, pp. 233–238, 2001.

[8] Sim, T., Kanade, T.: “Combining Models and Exemplars for Face Recognition: An Illuminating Example,” *Proceedings of the CVPR 2001 Workshop on Models versus Exemplars in Computer Vision*, December, 2001.

[9] Wiskott, L., Fellous, J.-M., Kruger, N., von der Malsburg, C.: “Face recognition by elastic bunch graph matching,” *IEEE Trans on PAMI*, Vol.19, pp.775–779, 1997.

[10] Liu, Y., Schmidt, K., Cohn, J., Weaver, R.L.: “Human facial asymmetry for expression-invariant facial identification,” *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FG’02)*, May, 2002.

[11] Kohonen, T.: *Self-Organizing Maps*, Springer-Verlag, 1995.

[12] Friedman, J.K., Tukey, J.W.: “A Projection Pursuit Algorithm for Exploratoty Data Analysis,” *IEEE Trans on computers*, vol. 23, pp. 881–889, 1974.

[13] Duda, R. O., Hart, P.E., Stork, D.G.: *Pattern Classification*, John Wiley & Sons, Inc., 2001.

[14] Bishop, C.M., Svensén, M.: “GTM: The Generative Topographic Mapping,” *Neural Computation*, Vol. 10. No. 1, pp. 215–234, 1998.

[15] Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.R.: “Fisher Discriminant Analysis with Kernels,” *IEEE Neural Networks for Signal Processing IX*, pp. 41–48, 1999.

[16] Carreira-Perpoñán, M.: “A Reivew of Dimension Reduction Techniques,” *Technical Report CS-96-09*, Dept. of Computer Science University of Sheffield, 1997.

[17] Jain, A., Duin, P., Mao, J.: “Statistical Pattern Recognition: A Review,” *IEEE Trans on PAMI*, vol. 22, No. 1, pp. 4–37, 2000.

[18] Pcekalska, E., Paclík, P., Duin, R. P.W.: “A Generalized Kernel Approach to Dissimilarity-based Classification,” *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2001.

[19] Chandrasiri, N. P., Park, M. C., Naemura, T., Harashima, H.: ”Personal Facial Expression Space based on Multidimensional Scaling for the Recognition Improvement”, Proc. IEEE ISSPA’99, pp. 943–946, 1999.

[20] Branson, K.: http://www-cse.ucsd.edu/∼kbranson/branson_isomap.pdf, 2002.

[21] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[22] http://www.uk.research.att.com/facedatabase.html

[23] Phillips, P.J., Moon, H.J., Rizvi, S.A., Rauss,. P.J.: “The FERET Evaluation Methodology for Face-Recognition Algorithms,” *IEEE Trans. on PAMI*, vol. 22, No. 10, pp. 1090–1104, 2000.

Real-World Application of Face Recognition over Time and Verification across Image Sequence for User Access Control

Gökçe Dane[1] and Matthew Barth[2]

[1]ECE Dept, University of California at San Diego, La Jolla, CA, 92093, USA
[2]EE Dept, University of California at Riverside, Riverside, CA, 92521, USA
gdane@ucsd.edu, barth@cert.ucr.edu

Abstract

Face recognition and verification is an important problem for many real-world tasks, such as user access control. In this paper, we describe and evaluate an automatic face recognition and verification (FRV) system that has been developed to support user access control for a shared-use vehicle system program which operates under real world conditions. In this application, three important FRV issues are discussed: 1) recognition of faces over time (i.e., months); 2) user verification using temporal image sequences; and 3) recognition across different kiosks. In order to perform robust recognition over time, a unique feature update method has been developed and implemented. Further, a method has been developed to select the best face image among an image sequence acquired in one vehicle-trip registration session for verification purposes. The implemented system has been operated for several months and carefully evaluated. Under real-world conditions, the proposed methods achieve 13% improvement in recognition and 15% improvement in verification compared to standard principal component analysis based techniques.

1. Introduction

Face recognition from still and video images is an important problem, which has many commercial and law enforcement applications [1]. Face recognition can be defined as the task of computing the similarity between two faces and matching a face with one or more subjects in a database. On the other hand, face verification (authentication) can be viewed as a one-to-one system that compares the biometric information presented by an individual with the biometric information stored in a database corresponding to that individual [2]. Although considerable progress has been made in the field of FRV [5-10], not many methods have been tested with data sets from real-world applications that have variable lighting and other conditions. Outside some exceptions (e.g., [3]), FRV algorithms are typically tested on a collection of a few hundred images, where the pictures are taken under well-controlled conditions.

In this work, the focus has been to develop and evaluate a face recognition and verification system by maintaining and updating a training set operating under real world conditions. The system supports user access control for vehicle registration as a part of a larger shared-use vehicle system program [13, 14]. In this program, subscribers utilize smartcards for multiple purposes. The goal is to identify the mismatches between smartcard-based user-IDs and the card users to prevent fraudulent usage. This is an important task not only for this particular project, but also for wider-scale applications such as automated teller machines (ATMs), building access, entry into secure areas, etc. In this framework, the FRV system is also used for user authentication and has been tested on a large data set of over 5000 images of approximately 100 people acquired over several months, where the collection of images consists of difficult recognition cases. The difficulty posed by this data set stems from the fact that the images are taken under different lighting conditions, at different times and locations, with different viewing of face directions and facial expression when the users perform their normal trip-registration process. For the face recognition stage, a feature update method has been developed to make it possible for the system to perform robust recognition over time. Further, for the face verification stage of the system, a new method has been developed that chooses the best face for verification and discards the rest of the images acquired in one trip registration session to improve the performance.

2. System description

The application domain of the developed face recognition/verification system is a shared-use vehicle system operating on the University of California-

Riverside campus called UCR IntelliShare. This intelligent car-sharing system allows multiple users to easily access a fleet of electric vehicles in order to improve mobility on campus. In this system, users utilize smartcards to gain access small kiosk buildings, check out vehicles with a touchscreen display, and then gain access to assigned vehicles. At the touchscreen kiosks in the station buildings, a user swipes his/her smartcard at the card reader to start the trip registration process. The user touches the screen to enter information such as anticipated destination, estimated trip distance, and number of occupants on the trip. Each time the user touches to screen to make a selection, a digital picture of the person is taken via a camera located at the top of the touchscreen kiosk. The image database for the FRV system is collected at two of five kiosks. The two registration kiosks with the camera systems are illustrated in Figure 1.

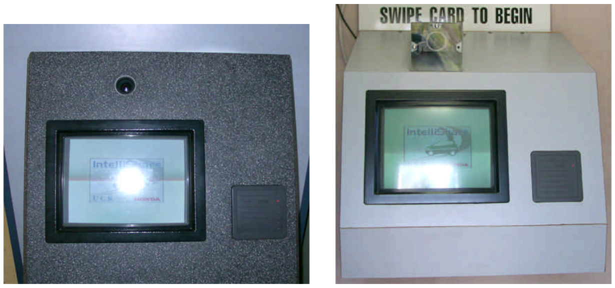


Figure 1. System touchscreen kiosks at two different locations used in the experiments

3. UCR IntelliShare face database

A large database of face images and/or image sequences is an important part of any FRV system. The content of such a database depends mostly on the purpose of the system. The UCR IntelliShare database is constructed by imaging the users via the cameras embedded in the touchscreen kiosks. The database has over 5000 images of 99 subjects (76 male and 23 female). The images of subjects were captured over three months at two separate kiosks (Figure 1). The acquisition of the images is performed with minimal cooperation from the users, as they only perform their typical trip registration process through the touchscreen display in the kiosks, which takes approximately 20 seconds. Since special instructions were not given to the users during the imaging process, the face pose and distance of the subjects to the camera varies greatly (e.g., see Figure 2). Therefore the conditions that the proposed FRV system operates are difficult to work with than those found in controlled laboratory conditions.

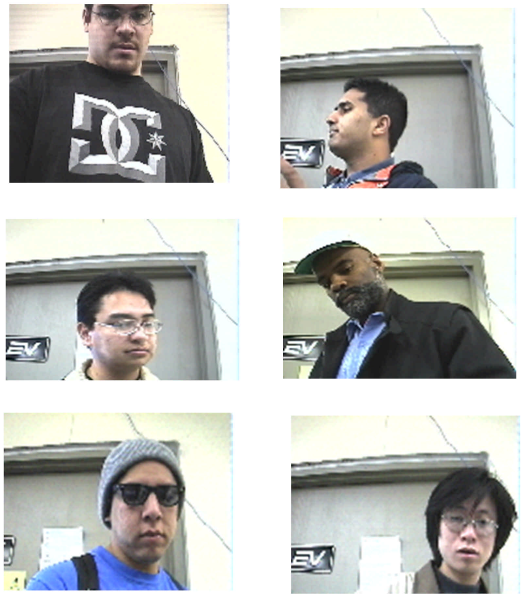


Figure 2. Example images from IntelliShare face database

4. Face recognition methodology

The entire FRV system is composed of three stages: *automatic face detection*, *face normalization*, and *feature extraction*. The overall effectiveness of the system strongly depends on the first stage of the system. In this stage, the face is extracted from the scene and approximate eye coordinates are located. For face detection, a combination of color and motion information is used which is followed by a template-based face search [15]. The motion change detection map in the face detection system is obtained by differencing two consecutive image frames, which are approximately 3 seconds apart on average. Simple differencing however cannot be used alone for locating faces due to noise, global illumination changes, and other moving objects in the scene. For this reason, skin color segmentation [5] is added as another cue for face location. The skin color model is obtained by training pixels from face regions, and applying a line fit model in normalized RGB color space to pixel color data. After obtaining a tighter face search area based on motion and color information, the face search map is downscaled to reduce the face search time. The search is carried out by using correlation over different-sized templates to compensate for size changes. After extracting the face from the refined search map, the primary facial marks are located and all the faces are normalized to a standard size to perform recognition and verification more efficiently.

The second stage is face normalization. The statistical approach that is used for face recognition requires a face-in-the-face-in-the-box model, where the extracted face is registered to the system in an 80x80 pixel box. To obtain the face-in-the-box model, each image is rotated automatically based on the eye centers found in the previous step. Accordingly, the line that passes through the central points of two eyes is kept horizontal. Then each face is normalized to a fixed scale to guarantee that the distance between the two eyes is kept constant to 40 pixels and each face fits in the same box. After face normalization, histogram equalization is performed for gray level normalization to partly reduce the effect of variable illumination strength.

The feature extraction stage of the proposed FRV system is based on the eigenface approach. This method, which was originally presented by [4], finds the principal components of the face image distribution or the eigenvectors of the covariance matrix of the set of face images. It is important to note that state of the art in face recognition has moved on since the eigenface approach, and many algorithms have been proposed [5-10]. Since our objective is to study training set maintenance, update over time and recognition across kiosks, we used the recognition algorithm given in [4] for the ease of its implementation.

To obtain recognition models, the proposed system goes into an off-line mode training stage. The training set used in this application, which has approximately 150 faces, is a subset of the larger UCR IntelliShare face database. The images in the training set are also acquired under uncontrolled conditions, and are manually selected. These 150 faces in the training set include at least one and at most three sample face of each subject. To obtain the basis vectors for recognition, principal component analysis is performed on the training set. Each subject's face is projected onto these basis vectors. The resulting coefficients with Euclidean distance measure are used as features in recognition and verification. Detailed explanation of these stages can be found in [15].

4.1. Updating features over time

The performance of face recognition algorithms degrades vastly over time even when tested with images that are taken under uncontrolled conditions [3]. To improve the performance over time, a feature update method is proposed as illustrated in Figure 3.

Let U be the unknown face to be tested, X be the feature vector of the unknown face and T be the feature vector of a known face in the training set. T_{ij} is the j^{th} feature vector of the i^{th} subject in the training database, where $1 \leq i \leq 99$, $1 \leq j \leq 3$. j varies between 1 and 3, since for some subjects more than one sample has been used in the training set. Let $p=E(X,T)$ be the Euclidean distance between X and T and let I_i be the updated feature vector. In the proposed method, for each day k of a certain week (in this application, the 8th week was used after the start of image database collection), and for each user i, the best match is selected based on p as follows: $p_{ik}=E(X_{ik},T_i)$ is calculated for all days (k with $1\leq k \leq 5$) of the updating period. Then $I_i = \min_k (p_{ik})$ is selected as the new feature vector. For subjects that has more than one instance (*i.e.* feature vector) in the training set the update is given as $I_i[j] = \min_k (p_{ik})$ where k^1 previous k found in $I_i[j-1], I_i[j-2], \ldots, I_i[1]$.

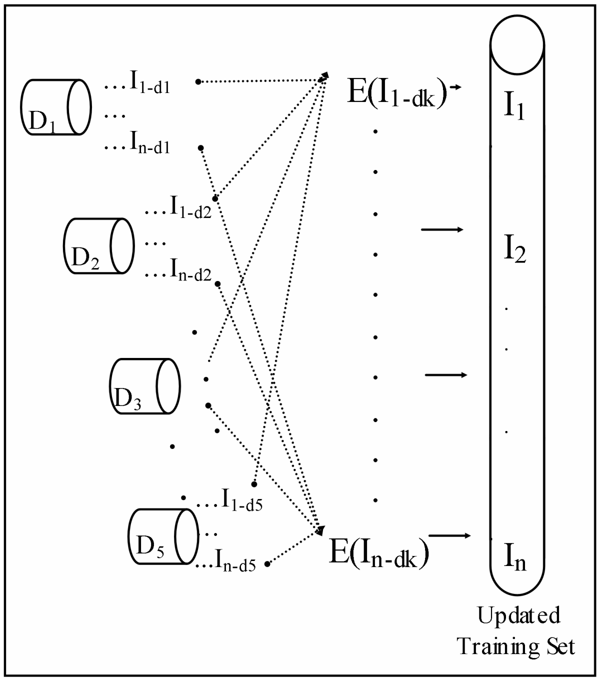


Figure 3 Feature Update Algorithm (D_k represents the image set acquired on Day k. I_{n-dk} represents the best match for subject n for k^{th} day. $E(I_{n-dk})$ selects the best match "I_n" for subject n for days from $k=1,\ldots,5$)

4.2. Face verification through an image sequence

Conventional methods perform face recognition and verification on single face images. To improve the verification performance, a method has been developed that uses image sequence as given in Figure 4. The system does not need to be trained by an image sequence. The same training set can be used for verification of both single and multiple images.

In this method, each image is considered as a classifier [16]. The distance measure p, which has been defined as a match score, is calculated for all images in the sequence acquired in one trip registration session. The one with the minimum score (*i.e.* max probability) is chosen as the best face for verification. If the test image is taken under different lighting conditions with different pose variances,

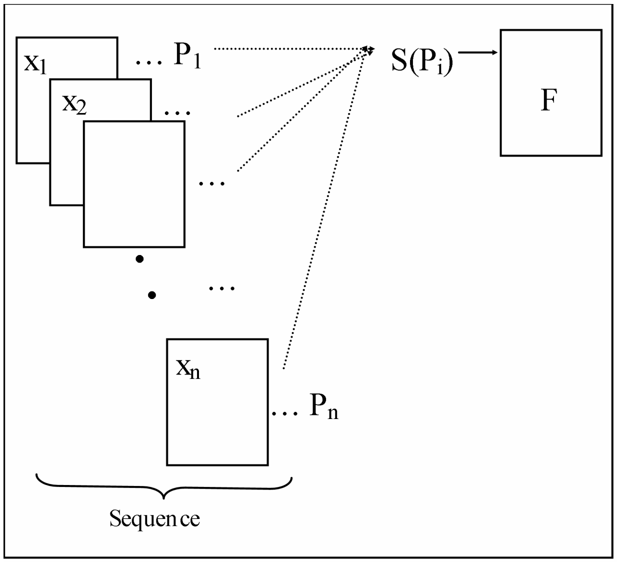


Figure 4. Combining image sequence for face verification (X_i represents the images in image sequence acquired for one subject in one session, P_i represents the match score of X_i. $S(P_i)$ selects the best face F to use for identity verification)

then the distance measure increases. Another reason for the greater distance measure is due to the failures in the automatic eye localization and normalization stage. If the eye centers cannot be located correctly and the face normalization is carried out inaccurately, the face recognition performance as well as verification performance degrades. However, by using the proposed method in verification, it is possible to compensate for some of the illumination and pose problems as well as possible incorrect automatic eye localization and normalization problems.

5. Test results and analysis

In the testing protocol, the testing and the training sets were separated, so that the images in both sets are distinct. The testing set (which includes more than 4000 images) is divided into separate subsets based on the time period that they have been acquired (e.g., March week1, May week 1, May week 2, etc.). Each subset has images taken on different days of the week, and even within each day there are many variations of the images of the same subject. The training set (which has about 150 images) is composed of the images which are acquired only in March week 1 and May week 1 (that is the first two weeks that the FRV system is up and running). The *closed universe* model was used for testing recognition performance, and the *open universe* model was used for testing verification performance [3]. In the closed model, every subject (i.e. every person) in the testing set is trained upon, but in the open universe model, the subject in the testing set may not have been trained upon and could be used as an impostor to test the verification performance.

5.1 Face recognition performance over time

In the first test, recognition performance degradation was analyzed over time. The rank *n* (Figure 5) represents the number of images that needs to be examined to get the desired level of performance. The statistics are given by percentage of correct identification as a function of rank. The horizontal axis represents the rank, and the vertical axis represents the percentage of correct matches. For the ease of visualization, the recognition results are given in terms of weekly time frames. When the recognition rates for each individual weekly set are examined, it is apparent that the performance degrades over time. The performance degrades from 52% to 36% in the best match and from 89% to 79% in the top 15 matches after a two months period.

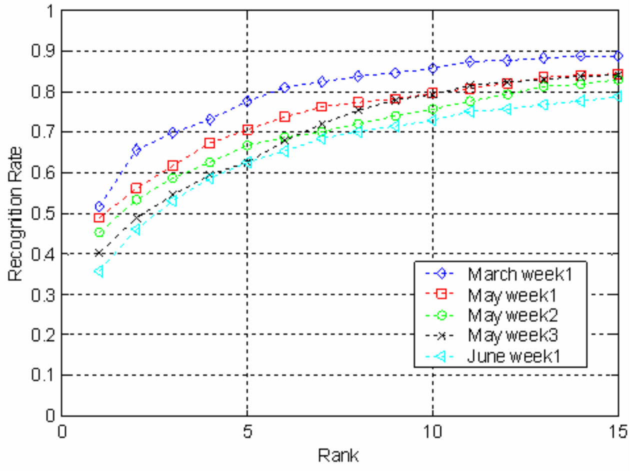


Figure 5. Face recognition performance in successive weeks

To illustrate the improvement obtained with proposed time-based feature update method, a recognition experiment was designed on an image set acquired during the tenth week after the start of system operation. The recognition experiment is performed by using both automatically and manually updated training sets obtained from the data set acquired in the ninth week. The difference between manual update and automatic update is that in the manual update method, the images of each subject in the database are chosen manually from the image data set, whereas in the automatic update method, the images of each subject used for training are chosen automatically by the proposed feature update algorithm. As can be seen from Figure 6, a 13% improvement can be obtained in the first match, and a 5% improvement in the top 15 matches. The improvement in the case of manual update is 5% more than that of the automatic update. The

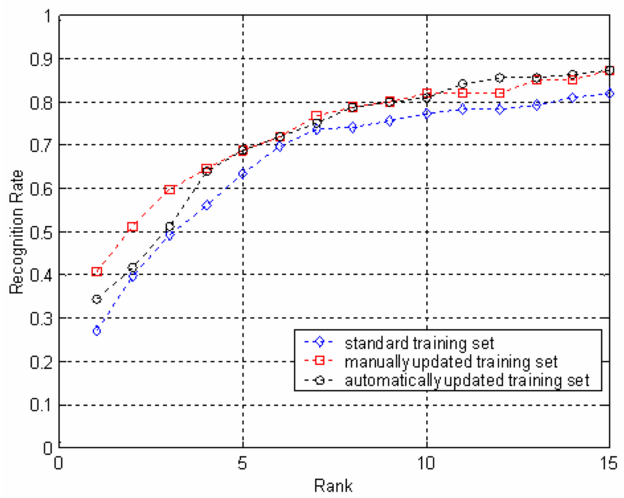


Figure 6. Face recognition performance after training set update

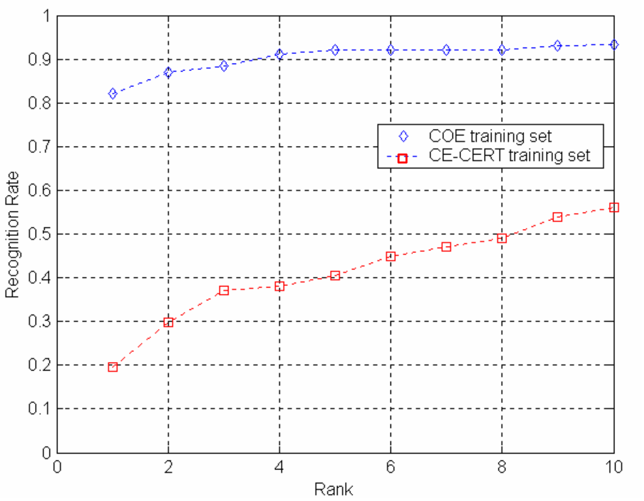


Figure 7. Face recognition performance in COE kiosk by using COE and CE-CERT training sets

reason is that in manual update, the images from the data set are chosen by considering the quality of image, such as being fully frontal, having acceptable face size, not having occlusion or uneven illumination, etc.

To ensure that the improvement is due to proposed technique, face detection results are checked manually. The face detection decision is given if eye coordinates are found correctly within a ±5 pixel range. The face detection performance stays same over time. For the same testing period, detection rates are as follows: March week 1= 94%, May week 1= 97%, May week 2= 89%, May week 3= 91%, June week1= 91%. So if we fix the detection rate, than the overall performance increase is due to proposed feature update method.

5.2 Face recognition performance across different locations

The purpose of this experiment is to show that the recognition performance depends strongly on the correlation of the training data to the testing data. In other words, if the recognition system is trained with images taken with the same physical set-up under consistent conditions with the testing data, superior recognition performance can be achieved.

To demonstrate this effect, we have conducted a recognition experiment by using two training sets (CE-CERT images and COE images) together with the manually labeled COE image data. The COE image set has different characteristics than the CE-CERT image and acquired with a different camera set-up. The best recognition performance of 81% on COE database is achieved by building the training set with the images from the same COE image data set as shown in Figure 7. On the other hand, if recognition tests are performed on COE data by using the training test built with CE-CERT image data, the recognition performance degrades to as low as 20% using the best match, and 56% using the first ten matches.

5.3 Face verification performance

The verification performance results are given by FRR (False Rejection Rate) and FAR (False Acceptance Rate) curves. The FRR is defined as the probability that a person is not authenticated to access the trip registration system even though you are the proper user. On the other hand, the FAR is the probability that someone other than the correct person is granted access to the system by using the person' s account or card. The performance of a verification system is judged by the Equal Error Rate (EER), which is the point in the Receiver Operating Characteristics (ROC) curve where FAR = FRR.

Figures 8 and 9 illustrate the ROC curves obtained by using standard verification method and the proposed method respectively. Both figures are obtained by using a data set of 274 images acquired in June week 1, from kiosk 1. For that specific set, 32 trip registration sessions were made. The False Rejection Rate is found by FR = EC/C, where EC is the number of client rejections, and C is the number of client claims. To find the False Acceptance Rate, impostors were introduced as follows: For each trip registration session, the actual user is excluded from the training set and given a false identity, where the false identity is found by a random number generated among the system users. Then the False Acceptance Rate is found by FA = EI/I, where EI is the number of impostor acceptances, and I is the number of Impostor claims.

When a comparison is made between the standard method where a single face shot is used for verification with the proposed method where an image sequence is used, it is apparent that EER falls from 30% to 15.4 %.

6. Conclusions

In this research, an automatic FRV system has been designed, implemented, and evaluated, operating under real world conditions. In particular, focus was placed on the problem of face recognition/verification over long periods of time by training set maintenance. A new method to update the feature space was introduced to make it possible for the system to perform robust recognition over time. Further, a verification strategy is described which uses image sequences. Using a temporal image sequence instead of a single image helps to overcome some of the problems such as pose, illumination, and incorrect face and eye coordinate location, which greatly affect the robustness of the performance. The effectiveness of both proposed strategies is demonstrated through experimental results. It has also been shown that recognition performance degrades vastly when a system is trained with face images taken in another location.

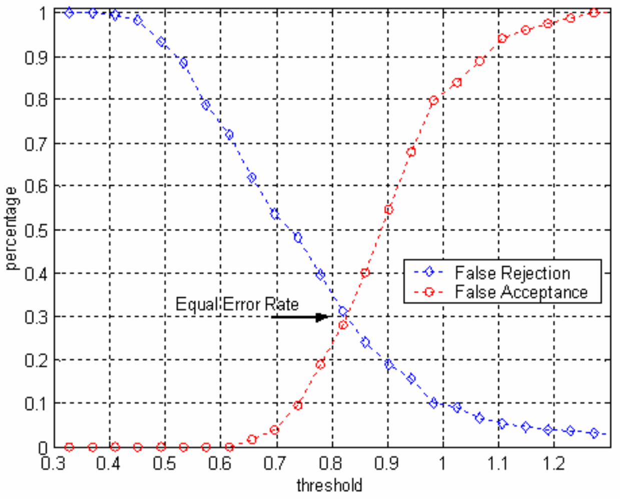


Figure 8. ROC curve for verification using single shot

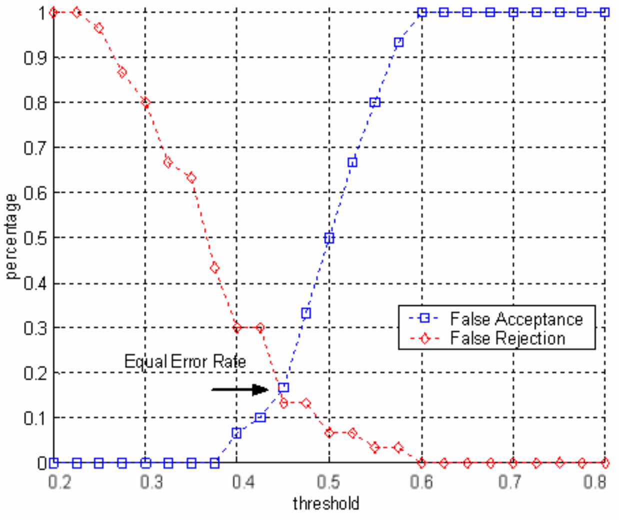


Figure 9. ROC curve for verification through image sequence

7. References

[1] R. Chellappa, C., L. Wilson, S. Sirohey, “Human and machine recognition of faces: A Survey”, *Proc. IEEE*, vol. 83, no. 5, 1995, pp. 705-740.

[2] C. L. Kotropoulos, A. Tefas, I. Pitas, “Frontal face authentication using discriminating grids with morphological feature vectors”, *IEEE Transactions on Multimedia*, vol. 2, no. 1, 2000, pp. 14-26.

[3] J., Phillips, H., Moon, P. Rauss, S., Rizvi, “The FERET evaulation methodology for face recognition algorithms”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no.10, 2000, pp. 1090-1103.

[4] M., A., Turk and A., P. Pentland, “Face recognition using eigenfaces”, *Proc. IEEE*, 1991.

[5] P., Belheumer, J., Hespana, D., Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection ", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 19, no.7, pp. 711-720, 1997.

[6] B., Moghaddam, W., Wahid, and A., Pentland, "Beyond eigenfaces: Probabilistic Matching for face recognition", - *IEEE*, 1998.

[7] J. Daugman, "Face and Gesture Recognition: Overview", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 19, no.7, pp. 675-676, 1997

[8] C., Liu, H., Wechsler, "Evolutionary pursuit and its application to face recognition", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, no. 6, pp. 570-582, 2000.

[9] Y. Adini, Y. Moses, S., Ullman, "Face recognition: The problem of compensating for changes in illumination direction", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 19, no. 7, pp. 721-732, 1997.

[10] J. Zhang, Y. Yan, M. Lades, "Face Recognition: Eigenfaces, elastic matching, and neural nets", *Proc. IEEE*, vol. 85, no. 9, pp. 1423-1435, September 1997.

[11] J. Terrillion, M., Shirazi, H., Fukomachi, S., Akamatsu, “Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images”, *Proc. IEEE*, 2000.

[12] Yamaguchi, O., Fukui, E., Maeda, K. “Face recognition using temporal image sequence”, *3rd IEEE Conf. of AFGR*, 1998, p.318-23. “ ”

[13] M. Barth, M. Todd, H. Murakami, “Using Intelligent Transportation System Technology in a Shared Electric Vehicle Program”, *Transportation Research Record*, No. 1731, pp. 88-95. Transportation Research Board, National Academy of Science, Washington D.C, 2000.

[14] M. Barth, M Todd, “Intelligent transportation system architecture for a multi-station shared vehicle system”, *IEEE ITSC*, 2000.

[15] G. Dane, “Real-world face recognition for detecting mismatched identities”, *MS Thesis*, University of California, Riverside, 2001.

[16] J. Kittler, M. Hatef, R. Duin, J. Matas, "On Combining Classifiers", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 20, no. 3, 1998.

A Model of Illumination Variation for Robust Face Recognition

Florent Perronnin and Jean-Luc Dugelay
Institut Eurécom
Multimedia Communications Department
BP 193, 06904 Sophia Antipolis Cedex, France
{*florent.perronnin, jean-luc.dugelay*}*@eurecom.fr*

Abstract

We recently introduced a novel approach to face recognition which consists in modeling the set of possible transformations between face images of the same person. While our previous work focused on geometric transformations to model facial expressions, in this article we consider feature transformations as a means to compensate for illumination variations. Although this approach requires to learn the set of possible illumination transformations through a training phase, we will show experimentally that the trained parameters are very robust. Even in the challenging case where the databases used to train the transformation model and to assess the performance of the system are very different, the proposed approach results in large improvements of the recognition rate.

1. Introduction

Pattern classification deals with the general problem of inferring classes from observations [1]. Hence, the success of a pattern classification system is based on its ability to distinguish between inter- and intra-class variabilities. Face recognition is a very challenging task as different faces have the same global shape while face images of the same person are subject to a wide range of variabilities including facial expressions, pose, illumination conditions, presence or absence of eyeglasses and facial hair, aging, occlusion, etc. Illumination, which will be the focus of this paper, remains one of the toughest variabilities to cope with as shown during the FERET evaluation [2] and the facial recognition vendor test 2000 [3].

It is possible to deal with the illumination at three different stages: during the *preprocessing*, the *feature extraction* or the *classification.*

Preprocessing algorithms for illumination compensation include general image processing tools such as histogram equalization and gamma correction [4]. A simple but very effective preprocessing, which is based on Weber's law, consists in applying a logarithm transform to the image intensity [5, 6]. Another class of preprocessing algorithms consists in separating an image into its reflectance and illumination fields [7]. An assumption which is generally made for this type of approach is that the luminance varies slowly across the image while sharp changes can occur in the reflectance.

At the feature extraction stage, the goal is to derive features that are invariant to illumination. Edge maps, derivatives of the gray level and Gabor features were compared in [5] and an empirical study showed that none of these features was sufficient to overcome the variations due to changes in the direction of illumination. Another idea is to *learn* features which are insensitive to illumination variations such as the Fisherfaces [8].

Finally, various algorithms have been proposed to cope with the illumination variation at the classification stage. The idea underlying [9] is that the set of images of an object in fixed pose, but under all possible illumination conditions, is a convex cone in the space of images that can be approximated by low dimensional linear subspaces. [10] proposed an approach based on 3D morphable models which encode both shape and texture information and an algorithm that recovers these parameters from a single face image.

We recently introduced a novel approach to face recognition which consists in modeling the set of possible transformations between face images of the same person [11]. While our previous work focused on *geometric* transformations to model facial expressions, we introduce in this article *feature* transformations as a means to compensate for illumination variations. This approach to illumination compensation, which works at the classification stage, involves a training phase to learn the set of possible illumination transformations. While approaches based on learning can suffer from poor generalization when the training and test sets are different, we will show experimentally the good generalization ability of our approach.

The remainder of this paper is organized as follows. A brief review of the probabilistic model of face transforma-

This work was supported in part by France Telecom Research & Development.

tion is given in the next section. Section 3 introduces our model of illumination transformation. Section 4 focuses on how to find jointly the best set of geometric and feature transformations between two face images. Finally, section 5 summarizes experimental results for a face identification task. While it is common to train and test a system on the same database, to assess the performance of our novel illumination compensation algorithm we used two very different databases. We think this is a much more realistic approach as, in practice, one never has access at training time to the exact test conditions. Even in this challenging case the proposed approach results in large improvements of the recognition rate.

2. A model of face transformation

2.1. Framework

While most face recognition techniques directly model the face, [11] models the set of possible *transformations* between face images of the same person. The global face transformation is approximated with a set of *local transformations* under the constraint that *neighboring* transformations must be consistent with each other.

Local transformations and consistency costs are embedded within the probabilistic framework of a 2D HMM. At any position on the query face image, the system is in one of a finite set of states where each state represents a local transformation. Emission probabilities model the cost of local transformations and transition probabilities relate states of neighboring regions and implement the consistency rules.

A major assumption in our system is that the intra-class variability is the same for all classes and, thus, that the model of face transformation is *shared* by all individuals. Hence, it can be trained on pairs of images of persons that are not enrolled in the system.

2.2. Local Transformations

Let us assume that we have two face images: a template image $\mathcal{F}_T$ and a query image $\mathcal{F}_Q$. Feature vectors are extracted on a sparse grid from $\mathcal{F}_Q$ and on a dense grid from $\mathcal{F}_T$. We then apply a set of local transformations at each position (i, j) of the sparse grid. In our previous work, these transformations were limited to geometric transformations and, more precisely, to translations. Each translation maps a feature vector of $\mathcal{F}_Q$ with a feature vector in $\mathcal{F}_T$.

Let $o_{i,j}$ be the observation extracted from $\mathcal{F}_Q$ at position (i, j) and let $q_{i,j}$ be the associated state (i.e. local deformation). If τ is a translation vector, the probability that at position (i, j) the system emits observation $o_{i,j}$, knowing that it is in state $q_{i,j} = \tau$, is $b^\tau_{i,j}(o_{i,j}) = P(o_{i,j}|q_{i,j} = \tau, \lambda)$ where $\lambda = (\lambda_T, \lambda_\mathcal{M})$. We separate λ into *face dependent* (FD) parameters λ_T which are extracted from $\mathcal{F}_T$ and *face independent transformation* (FIT) parameters $\lambda_\mathcal{M}$, i.e. the parameters of the shared transformation model $\mathcal{M}$. The emission probability $b^\tau_{i,j}(o_{i,j})$ represents the cost of matching $o_{i,j}$ with the corresponding feature vector in $\mathcal{F}_T$ that will be denoted $m^\tau_{i,j}$. $b^\tau_{i,j}(o_{i,j})$ is modeled with a mixture of Gaussians as linear combinations of Gaussians have the ability to approximate arbitrarily shaped densities:

$$b^\tau_{i,j}(o_{i,j}) = \sum_k w^k_{i,j} b^{\tau,k}_{i,j}(o_{i,j})$$

$b^{\tau,k}_{i,j}(o_{i,j})$'s are the component densities and the $w^k_{i,j}$'s are the mixture weights and must satisfy the following constraint: $\forall(i,j)$, $\sum_k w^k_{i,j} = 1$. Each component density is a D-variate Gaussian function of the form:

$$b^{\tau,k}_{i,j}(o_{i,j}) = \frac{\exp\left\{-\frac{1}{2}(o_{i,j} - \mu^{\tau,k}_{i,j})^T \Sigma^{k(-1)}_{i,j}(o_{i,j} - \mu^{\tau,k}_{i,j})\right\}}{(2\pi)^{\frac{N}{2}} |\Sigma^k_{i,j}|^{\frac{1}{2}}}$$

where $\mu^{\tau,k}_{i,j}$ and $\Sigma^k_{i,j}$ are respectively the mean and covariance matrix of the Gaussian, D is the size of feature vectors and $|.|$ is the determinant operator. We use a bi-partite model which separates the mean into additive FD and FIT parts:

$$\mu^{\tau,k}_{i,j} = m^\tau_{i,j} + \delta^k_{i,j} \tag{1}$$

where $m^\tau_{i,j}$ is the FD part of the mean. $w^k_{i,j}$, $\delta^k_{i,j}$ and $\Sigma^k_{i,j}$ are FIT parameters. Intuitively, $b^\tau_{i,j}$ should be approximately centered and maximum around $m^\tau_{i,j}$.

2.3. Neighborhood Consistency

The neighborhood consistency of the local transformations is ensured via the transition probabilities of the 2D HMM. We explain in the next section that a 2D HMM can be approximated by a set of interdependent horizontal and vertical 1D HMMs. The transition probabilities of the horizontal and vertical 1D HMMs are $P(q_{i,j} = \tau | q_{i,j-1} = \tau', \lambda)$ and $P(q_{i,j} = \tau | q_{i-1,j} = \tau', \lambda)$. They model respectively the horizontal and vertical elastic properties of the face at position (i, j) and are part of the face transformation model $\mathcal{M}$.

2.4. Turbo-HMMs

While HMMs have been extensively applied to 1D problems, the complexity of their extension to 2D grows exponentially with the data size and is intractable in most cases of interest. [12] introduced Turbo-HMMs (T-HMMs), in reference to the turbo error-correcting codes, to approximate the computationally intractable 2D HMMs. A T-HMM consists of horizontal and vertical 1D HMMs that "communicate" through an iterative process by inducing prior probabilities on each other. The T-HMM framework provides efficient formulas to 1) compute efficiently $P(\mathcal{F}_Q|\mathcal{F}_T, \mathcal{M})$, i.e. the probability that $\mathcal{F}_T$ and $\mathcal{F}_Q$ belong to the same person knowing the face transformation model $\mathcal{M}$, and 2) train automatically all the parameters of $\mathcal{M}$.

The computation of $P(\mathcal{F}_Q|\mathcal{F}_T, \mathcal{M})$ is based on a modified version of the forward-backward algorithm which is applied successively and iteratively on the horizontal and vertical 1D HMMs until they reach agreement.

The *Maximum Likelihood Estimation* (MLE) of the parameters of $\mathcal{M}$ is based on a modified version of the *Baum-Welch* algorithm. To train $\mathcal{M}$, we present pairs of pictures (a template and a query image) that belong to the same persons and optimize the transformation parameters $\lambda_{\mathcal{M}}$ to maximize the likelihood of the pairs of pictures.

3. Modeling the illumination variation

In this section, we will first show how to transform the illumination into an additive variability in the feature domain and then, how to constrain the illumination variation.

3.1. The illumination as an additive variability

The starting point of our approach is the well-known assumption that an image I can be seen as the product of a reflectance R and an illumination L [13]:

$$I(x, y) = R(x, y) \times L(x, y)$$

Applying the logarithm operator, we obtain:

$$\log I(x, y) = \log R(x, y) + \log L(x, y)$$

and the illumination turns into an additive term in the pixel domain. If the feature extraction involves only linear operators, such as the convolution, the illumination remains additive in the feature domain. Denoting F_d the linear feature extraction operator for the d-th dimension of the feature vectors and $o_{i,j} = \{o_{i,j}[1], ...o_{i,j}[D]\}$ the feature vector extracted at position (i, j), we get:

$$\begin{aligned} o_{i,j}[d] &= F_d\{\log I(x, y)\} \\ &= F_d\{\log R(x, y)\} + F_d\{\log L(x, y)\} \end{aligned}$$

Hence, if the illumination was constant in each feature component across the whole face, subtracting in each component the average value $\bar{o}[d]$ would be a simple approach to removing the undesired additive illumination term. However, the illumination is unlikely to be perfectly constant in each component. Moreover, when subtracting $\bar{o}[d]$, one may also discard useful reflectance information. Nevertheless, this simple combination of logarithm transform in the pixel domain and mean normalization in the feature domain, that will be referred to as the *Log-Mean Normalization* (or *LM-Norm*), and which, to the best of our knowledge, has never been suggested, will be tested in the section on experimental results.

Our goal is now to alleviate the unrealistic constraint of a constant illumination in each frequency band. As the system described in section 2 is designed to model additive variabilities, as expressed by equation (1), a first idea would be to train the Gaussian mixtures parameters, i.e. w's, δ's and Σ's, not only to model the facial expression variations, but also the various possible illumination conditions. Although this approach might first sound appealing, we believe it is suboptimal for two main reasons :

- A very large number of Gaussians would be necessary to model all the possible variabilities, increasing unreasonably the memory and CPU requirements.
- The choice of Gaussians at adjacent positions would be unconstrained, which is not satisfying as the illumination cannot vary in an arbitrary manner over the face.

However, the performance of this approach will also be evaluated in the section on experimental results and will serve as a baseline for our novel model of illumination transformation.

3.2. Constraining the illumination variation

The idea is to introduce feature transformations to model the illumination variation and to enforce consistency between feature transformations at adjacent positions in the same manner we enforced consistency between geometric transformations. Hence, our states which represent both local geometric and feature transformations are now doubly indexed: $q_{i,j} = (q^1_{i,j}, q^2_{i,j})$. $q^1_{i,j}$ is the geometric transformation part of the state and $q^2_{i,j}$ is the feature transformation part. If $q_{i,j} = (\tau, \phi)$, the emission probability $b^{\tau,\phi}_{i,j}$ is still modeled with a mixture of Gaussians:

$$b^{\tau,\phi}_{i,j} = \sum_k w^k_{i,j} b^{\tau,\phi,k}_{i,j}$$

where the $b^{\tau,\phi,k}_{i,j}$'s are D-variate Gaussians with means $\mu^{\tau,\phi,k}_{i,j}$ and covariance matrices $\Sigma^k_{i,j}$. The new means are of the form:

$$\mu^{\tau,\phi,k}_{i,j} = \mu^{\tau,k}_{i,j} + \phi = m^\tau_{i,j} + \delta^k_{i,j} + \phi$$

In [11] we only separated parameters into FD and FIT parameters. Here, we go one step further by separating the FIT parameters into geometrical transformation parameters and feature transformation parameters.

If we assume that geometric and feature transformations model respectively differences in facial expression and illumination between images, and that facial expression and illumination *variations* are mostly independent (i.e. a facial expression change between two adjacent positions has a limited impact on the illumination change between the same positions and vice versa), then the horizontal and vertical transition probabilities can be separated as follows:

$$\begin{aligned} P(q_{i,j}|q_{i,j-1}) &= P(q^1_{i,j}|q^1_{i,j-1}) \times P(q^2_{i,j}|q^2_{i,j-1}) \\ P(q_{i,j}|q_{i-1,j}) &= P(q^1_{i,j}|q^1_{i-1,j}) \times P(q^2_{i,j}|q^2_{i-1,j}) \end{aligned}$$

While the choice of a discrete number of geometric transformations is natural due to the discrete nature of the feature extraction grid of the template image, it is easier to deal with the illumination with an *infinite continuous* set of illumination states. We choose the horizontal and vertical illumination components of the transition probabilities to be D-variate Gaussians:

$$P(q^2_{i,j} = \phi|q^2_{i,j-1} = \phi') = P(q^2_{i,j} = \phi|q^2_{i-1,j} = \phi')$$
$$= \frac{\exp\left\{-\frac{1}{2}(\phi - \phi')^T S^{(-1)}(\phi - \phi')\right\}}{(2\pi)^{\frac{N}{2}}|S|^{\frac{1}{2}}}$$

In the following we will assume that the covariance matrix S is diagonal and therefore, that the components of the feature vectors are independent from each other. S is the only parameter of our illumination transformation model.

4. Finding the best transformation

Let $O = \{o_{i,j}\}$ and $Q = \{q_{i,j}\}$ denote respectively the set of all observations and states, with $i \in [1, I]$ and $j \in [1, J]$. Finding the best transformation between two face images requires to find the sequence of states Q^*, which satisfies:

$$Q^* = \arg\max_Q \log P(Q|O, \lambda) = \arg\max_Q \log P(O, Q|\lambda)$$

where $Q = (T, \Phi)$ and $T = \{\tau_{i,j}\}$ and $\Phi = \{\phi_{i,j}\}$ correspond respectively to the set of geometric and feature transformations. A central idea in our approach is to apply *iterative* passes to find *successively* the geometric and feature transformations that best explain the transformation between the two face images.

Let $Q_n = (T_n, \Phi_n)$ be the best set of states after the n-th iteration. Assuming for instance that we start by decoding geometric transformations, the steps of the algorithm are as follows:

1. Initialize Φ_0: $\forall(i, j)$, $\phi_{i,j} = 0$, i.e. we assume there is no illumination variation between the two images.

2. $T_n = \arg\max_T \log P(O, T|\Phi_{n-1}, \lambda)$, i.e. T_n maximizes the joint probability of observations and geometric transformations knowing Φ_{n-1}, the set of previously obtained feature transformations.

3. $\Phi_n = \arg\max_\Phi \log P(O, \Phi|T_n, \lambda)$, i.e. Φ_n maximizes the joint probability of observations and feature transformations knowing T_n, the set of geometric transformations previously obtained.

4. Go back to step 2 until T_n and Φ_n converge.

We will now detail the steps 2 and 3 of this algorithm.

4.1. Finding T_n

To find the best sequence of geometric transformations T_n, one applies the modified version of the forward-backward algorithm introduced in [12] and estimates the occupancy probabilities $\gamma_{i,j}(t) = P(q^1_{i,j} = t|O, \Phi_{n-1}, \lambda)$, i.e. the probability of being in state $q^1_{i,j} = t$ at position (i, j). At each position (i, j), we look for the best state τ:

$$\tau = \arg\max_t \gamma_{i,j}(t)$$

Although choosing the sequence of locally optimal states may not lead to the sequence of globally optimal states, this approximation is valid in the case where the best sequence of states accounts for most of the total probability.

If $\gamma_{i,j}(\tau, n)$ is the probability of being in state τ with the n-th mixture component accounting for $o_{i,j}$, the best Gaussian index k is given by:

$$k = \arg\max_n \gamma_{i,j}(\tau, n)$$

If τ and k are respectively the indexes of the best state and Gaussian at position (i, j), we introduce the quantity $\Psi^{\tau,k}_{i,j} = (o_{i,j} - \mu^{\tau,k}_{i,j})$ which can be interpreted as the variability that is left unexplained by the geometric transformations. Let $\Sigma^k_{i,j}$ be the covariance of the best Gaussian at position (i, j). In the following, for simplicity, we will drop the τ and k indexes and replace the notation $\Psi^{\tau,k}_{i,j}$ with $\Psi_{i,j}$ and $\Sigma^k_{i,j}$ with $\Sigma_{i,j}$.

4.2. Finding Φ_n

To find the best sequence of feature transformations Φ_n, we can pursue two different approaches: either apply directly the *Viterbi* algorithm, or a modified version of the *forward-backward*. In both cases, as $\Sigma_{i,j}$ and S the covariances of the emission and transition probabilities are assumed diagonal, it it simple to show that finding the best state sequence Φ can be done independently in each of the D dimensions. Therefore, if $\Psi_{i,j} = [\psi_{i,j}[1], ...\psi_{i,j}[D]]^T$, $\Sigma_{i,j} = \text{diag}\{\sigma_{i,j}[1]^2, ...\sigma_{i,j}[D]^2\}$ and $S = \text{diag}\{s[1]^2, ... s[D]^2\}$ in the following, we drop the dimension indexes and use the notations $\psi_{i,j}$, $\sigma^2_{i,j}$ and s^2.

4.2.1. Viterbi variant

We assume that transition probabilities are separable, i.e.:

$$P(q^2_{i,j}|q^2_{i-1,j}, q^2_{i,j-1}) \propto P(q^2_{i,j}|q^2_{i-1,j})P(q^2_{i,j}|, q^2_{i,j-1})$$

(see [12] for more details on this approximation). The joint likelihood $P(O, \Phi|T_n, \lambda)$ can be written as a product of emission probabilities and horizontal and vertical transition probabilities. For one given dimension, to find the best sequence of states Φ_n, we set $\partial \log P(O, \Phi|T_n, \lambda)/\partial\phi_{i,j} = 0$, $\forall(i, j)$ and obtain:

$$\begin{aligned} &\phi_{i-1,j} + \phi_{i+1,j} + \phi_{i,j-1} + \phi_{i,j+1} - \\ &\phi_{i,j}\left(\frac{s^2}{\sigma^2_{i,j}} + 4\right) = -\psi_{i,j}\left(\frac{s^2}{\sigma^2_{i,j}}\right) \ , \forall(i,j) \end{aligned}$$

with obvious modifications for $i = 1$ or I and $j = 1$ or J. This is a linear system of $I \times J$ equations with $I \times J$ unknowns. If equations are ordered properly, this system is banded with bandwidth $\min(I, J)$. Hence, the complexity of solving this system is in $\mathcal{O}((I \times J) \times \min(I, J)^2)$. We recall that there are D such systems to solve, one per dimension of the feature vectors.

At training time, to find the optimal s^2 which maximizes $\log P(O, \Phi|T_n, \lambda)$, we set $\partial \log P(O, \Phi|T_n, \lambda)/\partial s^2 = 0$ and obtain:

$$\hat{s}^2 = \frac{\sum_{i,j}\left[(\phi_{i,j} - \phi_{i-1,j})^2 + (\phi_{i,j} - \phi_{i,j-1})^2\right]}{(I-1) \times J + I \times (J-1)}$$

In the previous formula, s^2 is estimated with one pair of images. The extension to multiple pairs of images is straightforward.

4.2.2. Forward-backward variant

A complexity in $\mathcal{O}((I \times J) \times \min(I, J)^2)$ is much lower than the complexity of solving a general linear system of $I \times J$ equations with $I \times J$ unknowns which is in $\mathcal{O}((I \times J)^3)$. However it might still be too demanding if I and J are large. Therefore, we explored an alternative approach which is based on our modified forward-backward algorithm, as applied to T-HMMs [12]. The extension from discrete states HMMs to continuous states HMMs (also referred to as *state space models* or *SSMs*) consists mainly in replacing sums with integrals.

We define $\gamma_{i,j}(\phi) = P(q^2_{i,j} = \phi|O, T_n, \lambda)$, i.e. the probability of being in state ϕ at position (i, j). To find the states that best explain the illumination transformation, we choose the sequence of locally optimal states Φ, i.e.:

$$\phi_{i,j} = \arg\max_{\phi} \gamma_{i,j}(\phi)$$

We introduce the following vertical forward, backward and occupancy probabilities:

$$\begin{aligned} \alpha^{\mathcal{V}}_{i,j}(\phi) &= P(o_{1,j}, ...o_{i,j}, q^2_{i,j} = \phi|T_n, \lambda) \\ \beta^{\mathcal{V}}_{i,j}(\phi) &= P(o_{i+1,j}, ...o_{I,j}|q^2_{i,j} = \phi, T_n, \lambda) \\ \gamma^{\mathcal{V}}_{i,j}(\phi) &= P(q^2_{i,j} = \phi|o_{1,j}, ...o_{I,j}, T_n, \lambda) \end{aligned}$$

Defining the corresponding horizontal quantities is straightforward. As the emission and transition probabilities are Gaussians, if we initialize the occupancy probabilities γ's in a Gaussian manner, one can show that the forward, backward and occupancy probabilities are Gaussian shaped. The parameters of these Gaussians, i.e. their means and variances, will be respectively denoted $\mu^{\alpha\mathcal{V}}_{i,j}$, $\mu^{\beta\mathcal{V}}_{i,j}$, $\mu^{\gamma\mathcal{V}}_{i,j}$ and $\sigma^{\alpha\mathcal{V}^2}_{i,j}$, $\sigma^{\beta\mathcal{V}^2}_{i,j}$, $\sigma^{\gamma\mathcal{V}^2}_{i,j}$. It is easy to show that we have:

$$\mu^{\gamma\mathcal{V}}_{i,j} = \frac{\mu^{\alpha\mathcal{V}}_{i,j}\sigma^{\beta\mathcal{V}^2}_{i,j} + \mu^{\beta\mathcal{V}}_{i,j}\sigma^{\alpha\mathcal{V}^2}_{i,j}}{\sigma^{\alpha\mathcal{V}^2}_{i,j} + \sigma^{\beta\mathcal{V}^2}_{i,j}} \qquad \sigma^{\gamma\mathcal{V}^2}_{i,j} = \frac{\sigma^{\alpha\mathcal{V}^2}_{i,j}\sigma^{\beta\mathcal{V}^2}_{i,j}}{\sigma^{\alpha\mathcal{V}^2}_{i,j} + \sigma^{\beta\mathcal{V}^2}_{i,j}}$$

Successive horizontal and vertical passes of our modified forward-backward (extended to T-HMMs with an infinite continuous set of states) are applied iteratively to estimate $\mu^{\alpha\mathcal{V}}_{i,j}$, $\mu^{\beta\mathcal{V}}_{i,j}$, $\sigma^{\alpha\mathcal{V}^2}_{i,j}$ and $\sigma^{\beta\mathcal{V}^2}_{i,j}$ until convergence of the $\gamma^{\mathcal{H}}_{i,j}$ and $\gamma^{\mathcal{V}}_{i,j}$ probability densities. As we do not have access to $\gamma_{i,j}$ but to $\gamma^{\mathcal{H}}_{i,j}$ and $\gamma^{\mathcal{V}}_{i,j}$, a simple combination rule based on the minimum divergence criterion is to set:

$$\phi_{i,j} = \frac{\sigma^{\gamma\mathcal{V}^2}_{i,j}\mu^{\gamma\mathcal{H}}_{i,j} + \sigma^{\gamma\mathcal{H}^2}_{i,j}\mu^{\gamma\mathcal{V}}_{i,j}}{\sigma^{\gamma\mathcal{V}^2}_{i,j} + \sigma^{\gamma\mathcal{H}^2}_{i,j}}$$

The complexity of this algorithm is clearly in $\mathcal{O}(I \times J \times N)$ where N is the number of horizontal and vertical passes.

The optimal parameter s^2 is given by:

$$\hat{s}^2 = \frac{\sum_{i,j}\int_{\phi,\phi'}(\phi - \phi')^2\left[\xi^{\mathcal{H}}_{i,j}(\phi, \phi') + \xi^{\mathcal{V}}_{i,j}(\phi, \phi')\right]d\phi d\phi'}{(I-1) \times J + I \times (J-1)}$$

where $\xi^{\mathcal{H}}_{i,j}(\phi, \phi') = P(q^2_{i,j-1} = \phi, q^2_{i,j} = \phi'|O, T_n, \lambda)$ and $\xi^{\mathcal{V}}_{i,j}(\phi, \phi') = P(q^2_{i-1,j} = \phi, q^2_{i,j} = \phi'|O, T_n, \lambda)$. Introducing the notations $\rho^{\alpha\mathcal{H}}_{i,j} = s^2/(s^2 + \sigma^{\alpha\mathcal{H}^2}_{i,j})$ and $\rho^{\alpha\mathcal{V}}_{i,j} = s^2/(s^2 + \sigma^{\alpha\mathcal{V}^2}_{i,j})$, we get:

$$\begin{aligned} \hat{s}^2 \quad = \quad & \frac{\sum_{i,j}\left[(\mu^{\gamma\mathcal{H}}_{i,j} - \mu^{\gamma\mathcal{H}}_{i,j-1})^2 + (\mu^{\gamma\mathcal{V}}_{i,j} - \mu^{\gamma\mathcal{V}}_{i-1,j})^2\right]}{(I-1) \times J + I \times (J-1)} \\ + \quad & \frac{\sum_{i,j}\left[\rho^{\alpha\mathcal{H}}_{i,j-1}{\sigma^{\alpha\mathcal{H}}_{i,j-1}}^2 + {\rho^{\alpha\mathcal{H}}_{i,j-1}}^2\sigma^{\gamma\mathcal{H}^2}_{i,j}\right]}{(I-1) \times J + I \times (J-1)} \\ + \quad & \frac{\sum_{i,j}\left[\rho^{\alpha\mathcal{V}}_{i-1,j}{\sigma^{\alpha\mathcal{V}}_{i-1,j}}^2 + {\rho^{\alpha\mathcal{V}}_{i-1,j}}^2\sigma^{\gamma\mathcal{V}^2}_{i,j}\right]}{(I-1) \times J + I \times (J-1)} \end{aligned}$$

The term $(\mu^{\gamma\mathcal{H}}_{i,j} - \mu^{\gamma\mathcal{H}}_{i,j-1})^2 + (\mu^{\gamma\mathcal{V}}_{i,j} - \mu^{\gamma\mathcal{V}}_{i-1,j})^2$ corresponds to $(\phi_{i,j} - \phi_{i,j-1})^2 + (\phi_{i,j} - \phi_{i-1,j})^2$ in the re-estimation formula of the Viterbi variant (c.f. the previous section). The additional terms are due to the fact that the forward-backward algorithm integrates over all paths to estimate s^2 while Viterbi only takes into account the best path.

5. Experimental results

In this section, we will first introduce the databases used to train and test our system and briefly describe Gabor features. We will then evaluate the performance of the LM-Norm introduced in section 3.1 and finally the performance of our novel model of illumination transformation.

5.1. Databases

5.1.1. The FERET face database

To train our transformation model, we used the FERET face database [2]. 500 individuals were extracted from the FAFB set which contains frontal views that exhibit large variations in facial expressions but very little variability in terms of illumination. There are two images per person in the FAFB set. We also used the 200 individuals in the FAFC set which contains frontal views that exhibit large variations in illumination conditions and facial expressions. There are three images per person in the FAFC set. All the FERET images were pre-processed to extract 128x128 pixels normalized facial regions.

5.1.2. The YALE B face database

The YALE B face database [9] was used to assess the performance of our system. It contains the images of 10 subjects under 9 different poses and 64 illumination conditions. As the focus of this paper is on illumination compensation, we used only the set which contains frontal face images. We divided the database into the four traditional subsets $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$ according to the angle the light source makes with the axis of the camera (less than $12°$, between $12°$ and $25°$, between $25°$ and $50°$ and between $50°$ and $77°$). For each person, the 7 images in $\mathcal{S}_1$ were successively used as the enrollment image and the images in $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$ were used as test images which made a total of 26,600 comparisons. The same pre-processing that was applied to the FERET images was applied to the Yale B face images.

5.2. Gabor features

In our experiments, we used Gabor features which have long been successfully applied to face recognition and facial analysis. Assuming polar coordinates (ρ, θ), the spectral half plane is partitioned into M frequency and N orientation bands [14]:

$$G_{i,j}(\rho,\theta) = \exp\left\{-\frac{1}{2}\left[\frac{(\rho-\omega_{\rho_i})^2}{\sigma^2_{\rho_i}} + \frac{(\theta-\omega_{\theta_j})^2}{\sigma^2_{\theta_i}}\right]\right\}$$
$$\text{with } i \in [1,M] \text{ and } j \in [1,N]$$

The parameters ω_{ρ_i}, σ_{ρ_i}, ω_{θ_j} and σ_{θ_i} are defined as follows:

$$\omega_{\rho_i} = \omega_{min} + \sigma_0 \tfrac{(f+1)f^{i-1}-2}{f-1} \qquad \sigma_{\rho_i} = \sigma_0 f^{i-1}$$
$$\omega_{\theta_j} = \tfrac{(j-1)\pi}{N} \qquad \sigma_{\theta_i} = \tfrac{\pi\omega_{\rho_i}}{2N}$$

After preliminary experiments, we chose $\omega_{min} = \pi/24$, $\omega_{max} = \pi/3$, $f = \sqrt{2}$, $M = 4$ and $N = 6$, which resulted in 24 dimensional feature vectors. Gabor responses are obtained through the convolution of an image and the Gabor wavelets. We use the modulus of these responses as feature vectors which introduces a non-linearity in the computation of our features. Thus, the illumination cannot be considered as a perfectly additive term in the feature domain.

Feature vectors were extracted every 16 pixels of the query images and every 4 pixels of the template images in both horizontal and vertical directions.

5.3. Performance of the LM-Norm

The goal of this section is to assess the performance of the LM-Norm introduced in 3.1. In this first set of experiments, we applied straightforwardly the face transformation model introduced in [11] which does not make use of feature transformations.

When the LM-Norm is associated to Gabor features, the feature extraction consists of 3 steps:

1. logarithm transform in the pixel domain
2. Gabor features extraction
3. mean normalization in each frequency band

Gabor features combined with LM-Norm will be denoted *LM-GB* features. We compared the performance of these features to Gabor features that will be referred to as *GB* features and to features that combine steps 1 and 2 and that will be denoted *L-GB* features.

The face transformation model was trained on the FAFB data only. Hence, no information on illumination variations could be learned at training time. The transformation model was trained as described in [11] up to 8 Gaussians per mixture (Gpm). Figure 1 shows the results.

Averaging the performance over the 3 subsets, the identification rate is 68.0% for GB features compared to 74.0% for L-GB features and 84.8% for the LM-GB features. Note

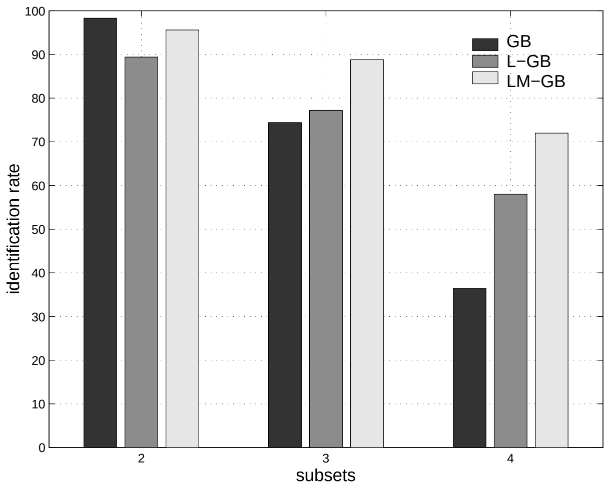


Fig. 1. Performance of GB (Gabor), L-GB (log + Gabor) and LM-GB (log + Gabor + mean normalization) features when the transformation model is trained solely on FAFB.

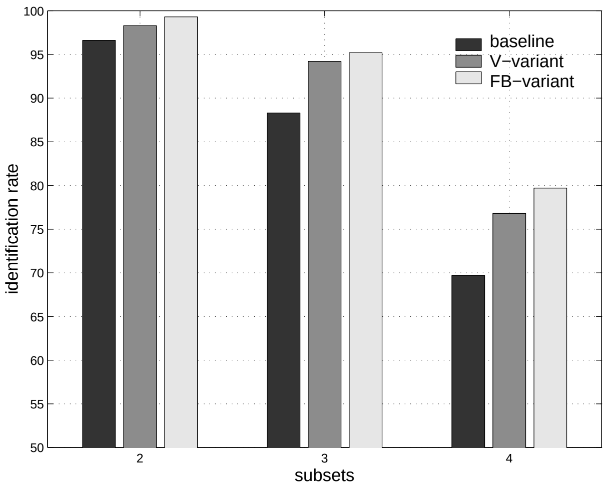


Fig. 2. Performance of the baseline system compared to the Viterbi and forward-backward variants (resp. V- and FB-variant) of our novel illumination compensation algorithm.

that with L-GB features the performance decreases significantly compared to GB features on the simple $\mathcal{S}_2$ subset which seems to indicate that the log transform has a negative impact on the recognition when there is little illumination variation.

We performed similar tests (not shown in this paper) with the GB, L-GB and LM-GB features on the popular Eigenfaces [15] and Fisherfaces [8] algorithms and observed similar trends. We would like to underline that, although we tested the combination of Gabor features and LM-Norm, we believe that LM-Norm could benefit to other "linear" features such as DCT features.

5.4. Performance of our novel approach

The goal of this second set of experiments is not only to assess the performance of our novel model of illumination transformation but also to assess the performance of the simple approach discussed in section 3.1, which is based solely on the transformation model introduced in [11] and which does not make use of any feature transformation. The latter algorithm will be referred to as the baseline.

For both algorithms, we applied a logarithm transform in the pixel domain prior to the extraction of Gabor features (L-GB features) as both methods require the illumination to be an additive term in the feature domain.

For our novel approach, we first trained our system up to 8 Gpm using only the FAFB data as explained in [11]. Then, using this model, we trained the covariance matrix S, which is the only parameter of the illumination transformation model, on the FAFC data only. The assumption is that, as the transformation model trained on FAFB already accounted for variations due to facial expressions, all the variability that remained unexplained was due to illumination. The diagonal elements of S were initialized to values close to 0 and then, 3 training iterations were applied. At both training and test time, the number of iterations of the decoding process described in section 4 was set to 3. To find Φ_n with the forward-backward variant of the algorithm described in section 4.2, we applied 5 horizontal and vertical passes.

For the baseline, we simply trained the system on both the FAFB and FAFC data up to 16 Gpm, instead of 8 Gpm, as more data was available.

Figure 2 shows the performance of the baseline compared to the Viterbi and forward-backward variants of our novel approach (resp. *V-variant* and *FB-variant*). Comparing Figures 1 and 2, one can see that adding the FAFC data increases on the average the identification rate of the baseline system from 74.0% to 84.1%. However, both variants of our novel approach clearly outperform the baseline, especially for the harder $\mathcal{S}_3$ and $\mathcal{S}_4$ subsets.

It is also interesting to notice that the FB-variant outperforms the V-variant. Actually, the latter one is optimal in the *Maximum-Likelihood* framework while our modified forward-backward based on the T-HMM framework is not guaranteed to be optimal. However, while Viterbi only takes into account the best path, i.e. the one that best explains the data, the forward-backward algorithm integrates over all paths. As explained in 4.2.2, this choice has an impact on the re-estimation of S and we believe that the difference in performance is mainly due to the difference in the re-estimation formula. The average identification rate of the V-variant and FB-variant over the three subsets are respectively 89.1% and 90.8%.

We also compared our novel approach with the eigenfaces [15] and Fisherfaces [8]. Especially Fisherfaces were shown to compensate for illumination variations if trained with the appropriate data. To carry out a fair comparison, we did not apply these algorithms directly on the gray level images but on their LM-GB representations. A feature vector was extracted every four pixels of the images in both horizontal and vertical directions. The eigen- and Fisher-spaces were trained on the FAFB and FAFC sets as was done for our baseline system. The best identification rates we obtained for eigenfaces and Fisherfaces are respectively 87.1% and 83.1%. The fact that eigenfaces outperform Fisherfaces is not surprising considering the small number of training observations per class and the mismatch between training and test conditions [16].

Finally, we would like to stress the fact that our novel algorithm is very efficient as it takes on the average to our best system less than 25 ms to compare two images on a 2 GHz Pentium 4 with 1 GB RAM.

6. Conclusion and future work

In this paper, we introduced a novel approach to illumination compensation, which consists in modeling the set of possible illumination transformations between face images of the same person. This approach is naturally embedded in a face recognition system which already models transformations between face images due to facial expressions. We showed experimentally that, even in the challenging case where we trained and tested our system on two different databases, our novel approach to illumination compensation resulted in large improvements of the recognition rate. Note that our results are competitive with state of the art results recently published on the YALE B database [7].

However, much work remains to be done to perfectly compensate for illumination variations. For the challenging $\mathcal{S}_4$ subset, the best identification rate we obtain is close to 80%. Although this corresponds to an almost 70% relative error rate reduction compared to the same system without any illumination compensation, we are still far from the almost perfect recognition rate we get for the simpler $\mathcal{S}_2$ subset. We believe that one limitation of our current approach is the fact that the covariance matrix S in our illumination transformation model is fixed for all pairs of images. We think that S should incorporate both some a priori knowledge learned off-line through a training phase, as is currently the case, but also some information which is dependent on the pairs of images that need to be compared.

Finally, we would like to point out that, while our model of illumination compensation has been introduced in the context of face recognition, it could benefit to other research areas. As our original approach to face recognition has a lot in common with motion estimation algorithms, and especially MAP estimation of dense motion [4], we think that our approach could be applied to the difficult problem of motion estimation in the presence of illumination variations.

7. References

[1] J. Schürmann, *Pattern classification, a unified view of statistical and neural approaches*, John Wiley & Sons, Inc., 1996.

[2] P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss, “The feret evaluation methodology for face recognition algorithms,” *IEEE Trans. on PAMI*, vol. 22, no. 10, pp. 1090–1104, Oct 2000.

[3] D. M. Blackburn, M. Bone and P. J. Phillips, “Face recognition vendor test 2000: evaluation report,” Tech. Rep., 2001.

[4] A. Bovik, *Handbook of image and video processing*, Academic Press, 2000.

[5] Y. Adini, Y. Moses and S. Ullman, “Face recognition: the problem of compensating for changes in illumination direction,” *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 721–732, July 1997.

[6] M. Savvides and V. Kumar, “Illumination normalization using logarithm transforms for face authentication,” in *IAPR AVBPA*, 2003, pp. 549–556.

[7] R. Gross and V. Brajovic, “An image preprocessing algorithm for illumination invariant face recognition,” in *IAPR AVBPA*, 2003, pp. 10–18.

[8] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Trans. on PAMI*, vol. 19, no. 7, pp. 549–556, July 1997.

[9] 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. on PAMI*, vol. 23, no. 6, pp. 643–660, June 2001.

[10] V. Blanz, S. Romdhani and T. Vetter, “Face identification across different poses and illuminations with a 3d morphable model,” in *IEEE AFGR*, 2002.

[11] F. Perronnin, J.-L. Dugelay and K. Rose, “Deformable face mapping for person identification,” in *IEEE ICIP*, 2003.

[12] F. Perronnin, J.-L. Dugelay and K. Rose, “Iterative decoding of two-dimensional hidden markov models,” in *IEEE ICASSP*, 2003, vol. 3, pp. 329–332.

[13] B. K. P. Horn, *Robot Vision*, Mc Graw-Hill, New-York, 1986.

[14] B. Duc, S. Fischer and J. Bigün, “Face authentication with gabor information on deformable graphs,” *IEEE Trans. on PAMI*, vol. 8, no. 4, pp. 504–516, April 1999.

[15] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *IEEE CVPR*, 1991, pp. 586–591.

[16] A. M. Martìnez and A. C. Kak, “Pca versus lda,” *IEEE Trans. on PAMI*, vol. 23, no. 2, pp. 228–233, Feb 2001.

Augmenting Frontal Face Models for Non-Frontal Verification

Conrad Sanderson and Samy Bengio
IDIAP, Rue du Simplon 4, CH-1920 Martigny, Switzerland
conradsand@ieee.org, bengio@idiap.ch

Abstract

In this work we propose to address the problem of non-frontal face verification when only a frontal training image is available (e.g. a passport photograph) by augmenting a client's frontal face model with artificially synthesized models for non-frontal views. In the framework of a Gaussian Mixture Model (GMM) based classifier, two techniques are proposed for the synthesis: UBMdiff and LinReg. Both techniques rely on a priori information and learn how face models for the frontal view are related to face models at a non-frontal view. The synthesis and augmentation approach is evaluated by applying it to two face verification systems: Principal Component Analysis (PCA) based and DCTmod2 [29] based; the two systems are a representation of holistic and non-holistic approaches, respectively. Results from experiments on the FERET database suggest that in almost all cases, frontal model augmentation has beneficial effects for both systems; they also suggest that the LinReg technique (which is based on multivariate regression of classifier parameters) is more suited to the PCA based system and that the UBMdiff technique (which is based on differences between two general face models) is more suited to the DCTmod2 based system. The results also support the view that the standard DCTmod2/GMM system (trained on frontal faces) is less affected by out-of-plane rotations than the corresponding PCA/GMM system; moreover, the DCTmod2/GMM system using augmented models is, in almost all cases, more robust than the corresponding PCA/GMM system.

1. Introduction

In the context of *frontal* faces, recent approaches to face recognition (here we mean both identification and verification) are able to achieve very low error rates (e.g. [19]). A more realistic and challenging task is to verify a face at a non-frontal view when only one (frontal) training image is available (e.g. a passport photograph).

While the task of view-independent recognition has been addressed through the use of training images (for the person to be recognized) at multiple views (e.g. [22]), the much harder task of using only one training image has received relatively little attention (e.g. [2, 21]). Whereas it is possible to use 3D approaches to address the single training image problem (e.g. [1, 17]), here we concentrate on extending two well understood 2D based techniques. In particular, we will extend the Principal Component Analysis (PCA) based approach [30] and the recently proposed DCTmod2 based approach [29]. In both cases we employ a Gaussian Mixture Model (GMM) based classifier [25], which is central to our extensions.

The PCA/GMM system is an extreme example of a holistic system where the spatial relation between face characteristics (such as the eyes and nose) is rigidly kept (with the advantage of robustness to compression artefacts & additive noise [28]). Conversely, the DCTmod2/GMM approach is an extreme example of a non-holistic approach; here, the spatial relation between face characteristics is effectively lost (which results in robustness to translations [4]). In between the two extremes are systems based on multiple template matching [3], modular PCA [20, 22], Pseudo 2D Hidden Markov Models (HMMs) [10, 26] and heuristic approaches such as Elastic Graph Matching (EGM) [8, 16].

Generally speaking, an appearance based face recognition system can be thought of as being comprised of:

1. Face localization and segmentation
2. Normalization
3. Feature extraction
4. Classification

The second stage (normalization) usually involves an affine transformation (to correct for size and rotation), but it can also involve an illumination normalization (however, illumination normalization may not be necessary if the feature extraction method is robust). In this paper we shall concentrate on the last stage (and thus postulate that the preceding steps have been performed correctly).

Some approaches to addressing the single training image problem involve the synthesis of new face images (at various angles) based on *a priori* information (e.g. [2, 21]). In these approaches, the image synthesis comes before the usual step of feature extraction. A question thus arises: if we are only interested in recognition and hence we are going to extract features from synthesized images, why not synthesize the features instead? If we follow this line of thinking, a natural followup question is: instead of synthesizing features with which we are going to train a classifier, why not directly synthesize the classifier's parameters? This is in fact the central idea of our proposed extensions, sketched below.

Using *a priori* information in the form of a set of faces at different views (these faces will never be used during performance evaluation), we construct face models for specific views (by "model" we mean a GMM); we then find the *differences* between the model for the frontal view and, say, the model for the +25^o view. Let us now suppose that we wish to enroll a new client in our face verification system and we only have their frontal view; given a face model created from their frontal view, we can synthesize a model for +25^o by applying the *a priori* differences to the client's frontal model. In order for the system to automatically handle the two views, we then augment the client's frontal model by concatenating it with the newly synthesized +25^o model. We can of course repeat this procedure for other views.

The proposed synthesis and augmentation approach thus differs from the approach presented in [2, 21] where actual face images for non-frontal views were synthesized; the synthesized images shown in [2] have considerable artefacts, which we believe can easily lead to a decrease in performance. The proposed approach is somewhat related to [18] where a feature transformation approach is employed in the context of an EGM based classifier. We note that in [18] manual intervention is required, while our proposed approach is auto-

The authors thank the Swiss National Science Foundation for supporting this work through the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2). The authors also thank Andrzej Drygajlo, Daniel Gatica-Perez, Sebastien Marcel, Alexei Pozdnoukhov and Alessandro Vinciarelli for helpful suggestions.

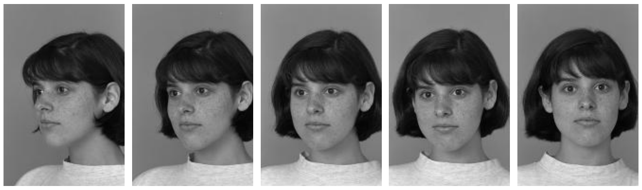


Fig. 1. Images of subject 00647 from the FERET database for (from left to right) -60^o, -40^o, -25^o, -15^o and 0^o views; note that the angles are approximate.

matic; moreover, unlike [18], our approach is based on a statistical framework. The augmentation part of our proposed approach is related to [14]; the main difference being that in [14] features from the client's many *real* images are used to extend the client's face model, while in our proposed approach we synthesize the models to represent the face of a client at various non-frontal angles, without having access to the client's real images.

The rest of the paper is organized as follows. In Section 2 we briefly describe the database used in the experiments and the pre-processing of the images. In Sections 3 and 4 we overview the DCTmod2 and PCA based feature extraction techniques, respectively. Section 5 provides a concise description of the GMM based classifier and the different training strategies used when dealing with DCTmod2 and PCA based features. In Section 6 we describe two techniques used for synthesizing non-frontal models as well as a method to address the problem of correspondence between two GMMs. Section 7 details the process of concatenating two or more GMMs. Section 8 is devoted to experiments evaluating the two synthesis techniques and the use of augmented models. The paper is concluded and future work is suggested in Section 9.

2. FERET Database: Setup & Pre-Processing

In our experiments we utilized face images from the FERET database [23]. In particular, we used images from the *ba*, *bb*, *bc*, *bd*, *be*, *bf*, *bg*, *bh* and *bi* subsets, which represent views of 200 persons for (approximately) 0^o (frontal), +60^o, +40^o, +25^o, +15^o, -15^o, -25^o, -40^o and -60^o, respectively; thus for each person there are nine images. Example images are shown in Fig. 1.

The 200 persons were split into three disjoint groups: group A, group B and impostor group; the impostor group is comprised of 20 persons, resulting in 90 persons in groups A and B. Throughout the experiments, group A is used as a source of *a priori* information while the impostor group and group B are used for verification tests (i.e. clients come from group B). Thus in each verification trial there is 90 true claimant accesses and 90×20=1800 impostor attacks; moreover, in each verification trial the view of impostor faces matched the testing view.

In order to reduce the effects of variations possible in real life (such as facial expressions, hair styles, clothes and ornaments) closely cropped faces are used instead of full face images [5]. In particular, we used the location of the eyes to normalize the inter-ocular distance and extract a 56×64 (rows × columns) face window containing the area from the eyebrows to the nose (inclusive). Example face windows are shown in Fig. 2.

Since in this paper we are proposing extensions to existing 2D approaches, we obtain normalized face windows for non-frontal views exactly in the same way as for the frontal view; this has a significant side effect: for large deviations from the frontal view (such as -60^o and +60^o) the effective size of facial characteristics is significantly larger than for the frontal view. The non-frontal face windows thus differ from the frontal face windows not only in terms of out-of-plane rotation of the face, but also scale.

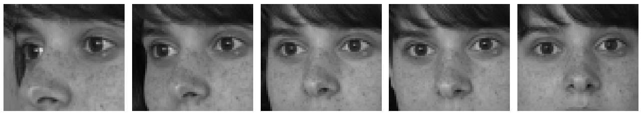


Fig. 2. Extracted face windows from images in Fig. 1.

Overlap (N_O)	Vectors (N_V)	Spatial width
0	30	24
1	35	22
2	56	20
3	80	18
4	143	16
5	255	14
6	621	12
7	2585	10

Table 1. Number of DCTmod2 feature vectors extracted from a 56×64 face using N_P=8 and varying overlap; also shows the effective spatial width (& height) in pixels for each feature vector.

3. Feature Extraction: DCTmod2 Based Sys.

In DCTmod2 feature extraction [29] a given face image is analyzed on a block by block basis; each block is $N_P \times N_P$ (here we use N_P=8) and overlaps neighboring blocks by N_O pixels. Each block is decomposed in terms of 2D Discrete Cosine Transform (DCT) basis functions [13]. A feature vector for each block is then constructed as:

$$\vec{x} = \left[\; \Delta^h c_0 \;\; \Delta^v c_0 \;\; \Delta^h c_1 \;\; \Delta^v c_1 \;\; \Delta^h c_2 \;\; \Delta^v c_2 \;\; c_3 \;\; c_4 \; \ldots \; c_{M-1} \; \right]^T \quad (1)$$

where c_n represents the n-th DCT coefficient, while $\Delta^h c_n$ and $\Delta^v c_n$ represent the horizontal & vertical delta coefficients respectively; the deltas are computed using DCT coefficients extracted from neighboring blocks. Compared to traditional DCT feature extraction [10], the first three DCT coefficients are replaced by their respective horizontal and vertical deltas in order to reduce the effects of illumination changes, without losing discriminative information. In this study we use M=15 (choice based on [29]), resulting in an 18 dimensional feature vector for each block.

The degree of overlap (N_O) has two effects: the first is that as overlap is increased the spatial area used to derive one feature vector is decreased; the second is that as the overlap is increased the number of feature vectors extracted from an image grows in a quadratic manner. Table 1 shows the amount of feature vectors extracted from 56×64 face using our implementation of the DCTmod2 extractor.

As will be shown later, the larger the overlap (and hence the smaller the spatial area for each feature vector), the more the system is robust to out-of-plane rotations.

4. Feature Extraction: PCA Based System

In PCA based feature extraction [30], a given face image is represented by a matrix containing grey level pixel values; the matrix is then converted to a face vector, $\vec{f}$, by concatenating all the columns; a D-dimensional feature vector, $\vec{x}$, is then obtained by:

$$\vec{x} = \mathbf{U}^T (\vec{f} - \vec{f}_\mu) \quad (2)$$

where $\mathbf{U}$ contains D eigenvectors (corresponding to the D largest eigenvalues) of the training data covariance matrix, and $\vec{f}_\mu$ is the mean of training face vectors. In our experiments we use frontal faces from group A to find $\mathbf{U}$ and $\vec{f}_\mu$. If robustness to illumination changes is required, an extension such as *enhanced PCA* can be utilized [28].

It must be emphasized that in the PCA based approach, one feature vector represents the entire face, while in the DCTmod2 approach one feature vector represents only a small portion of the face.

5. GMM Based Classifier

The distribution of training feature vectors for each person is modeled by a GMM. Given a claim for client C's identity and a set of (test) feature vectors $X = \{\vec{x}_i\}_{i=1}^{N_V}$ supporting the claim, the average log likelihood of the claimant being the true claimant is found with:

$$\mathcal{L}(X|\lambda_C) = \frac{1}{N_V} \sum\nolimits_{i=1}^{N_V} \log p(\vec{x}_i|\lambda_C) \tag{3}$$

where:

$$p(\vec{x}|\lambda) \;=\; \sum\nolimits_{j=1}^{N_G} w_j \, \mathcal{N}(\vec{x}; \vec{\mu}_j, \mathbf{\Sigma}_j) \tag{4}$$

$$\lambda \;=\; \{w_j, \vec{\mu}_j, \mathbf{\Sigma}_j\}_{j=1}^{N_G} \tag{5}$$

Here, $\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma})$ is a D-dimensional Gaussian function with mean $\vec{\mu}$ and diagonal covariance matrix $\mathbf{\Sigma}$:

$$\mathcal{N}(\vec{x}; \vec{\mu}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(\vec{x}-\vec{\mu})^T \mathbf{\Sigma}^{-1} (\vec{x}-\vec{\mu})\right) \tag{6}$$

λ_C is the parameter set for client C, N_G is the number of Gaussians and w_j is the weight for Gaussian j (with constraints $\sum_{j=1}^{N_G} w_j = 1$ and $\forall\, j : w_j \geq 0$).

Given the average log likelihood of the claimant being an impostor, $\mathcal{L}(X|\lambda_{\overline{C}})$, an opinion on the claim is found using:

$$\Lambda(X) = \mathcal{L}(X|\lambda_C) - \mathcal{L}(X|\lambda_{\overline{C}}) \tag{7}$$

The verification decision is reached as follows: given a threshold t, the claim is accepted when $\Lambda(X) \geq t$ and rejected when $\Lambda(X) < t$. In our experiments we use a global threshold to obtain performance as close as possible to the Equal Error Rate (EER) (i.e. where the false rejection rate is equal to the false acceptance rate), following the popular practice used in the speaker verification field [7, 11].

Methods for obtaining the parameter set for the impostor model ($\lambda_{\overline{C}}$) and each client are described in the following sections.

5.1. Classifier Training: DCTmod2 Based System

First, a Universal Background Model (UBM) is trained with a form of the Expectation Maximization (EM) algorithm [6, 9] using *all* 0^o data from group A; here the EM algorithm tunes the model parameters to optimize the Maximum Likelihood (ML) criterion (i.e. so that the likelihood of the training data is maximized).

The parameters (λ) for each client model are then found by using the client's training data and adapting the UBM (the number of Gaussians is varied in the experiments); the adaptation is accomplished using a different form of the EM algorithm, often referred to as maximum *a posteriori* (MAP) estimation [12, 25]. The two instances of the EM algorithm are summarized in appendixes A and B.

Since the UBM is a good representation of a general face, it is also used to find the likelihood of the claimant being an impostor, i.e.:

$$\mathcal{L}(X|\lambda_{\overline{C}}) = \mathcal{L}(X|\lambda_{ubm}) \tag{8}$$

5.2. Classifier Training: PCA Based System

The image subset from the FERET database that is utilized in this work has only one frontal image per person; in PCA-based feature extraction, this results in only one training vector, leading to necessary constraints in the structure of the classifier and the classifier's training paradigm.

The UBM and all client models (for frontal faces) are constrained to have only one component (i.e. one Gaussian). As for the DCTmod2 system (described above), the parameters for the UBM are found by running the EM algorithm on all data from group A. Instead of MAP estimation, each client model inherits the covariance matrix from the UBM; moreover, the mean of each client model is taken to be the single training vector for that client.

6. Synthesizing Models for Non-Frontal Views

6.1. UBMdiff Technique

Let us suppose that we have two UBMs, $\lambda_{ubm}^{0^o}$ and $\lambda_{ubm}^{+25^o}$ (trained using *a priori* data) that describe a general face for a view at 0^o and +25^o, respectively. Let us define the set of parameters which describes the difference between the two UBMs as:

$$\Delta^{+25^o} = \left\{ \, w_{\Delta,i}^{+25^o} \, , \, \vec{\mu}_{\Delta,i}^{+25^o} \, , \, \vec{\sigma}_{\Delta,i}^{+25^o} \, \right\}_{i=1}^{N_G} \tag{9}$$

The parameters are defined as:

$$w_{\Delta,i}^{+25^o} = w_{ubm,i}^{+25^o} / w_{ubm,i}^{0^o} \tag{10}$$

$$\vec{\mu}_{\Delta,i}^{+25^o} = \vec{\mu}_{ubm,i}^{+25^o} - \vec{\mu}_{ubm,i}^{0^o} \tag{11}$$

$$\left(\vec{\sigma}_{\Delta,i}^{+25^o}\right)^T = [\ \sigma_{\Delta,i,d}\ ]_{d=1}^{D} = \left[\mathbf{\Sigma}_{ubm,i,(d,d)}^{+25^o} / \mathbf{\Sigma}_{ubm,i,(d,d)}^{0^o}\right]_{d=1}^{D} \tag{12}$$

where $\mathbf{\Sigma}_{ubm,i,(d,d)}^{+25^o}$ denotes the element at row d and column d (i.e. d-th diagonal) of $\mathbf{\Sigma}_{ubm,i}^{+25^o}$. Since the two UBMs are a good representation of a general face at the two views, and each client model is derived from the 0^o UBM, it is reasonable to assume that we can apply the above difference to client C's 0^o model to synthesize a +25^o model. Formally, the parameters for the +25^o model are:

$$\lambda_C^{+25^o} = \left\{ w_{C,i}^{+25^o} \, , \vec{\mu}_{C,i}^{+25^o} \, , \mathbf{\Sigma}_{C,i}^{+25^o} \right\}_{i=1}^{N_G} \tag{13}$$

and are synthesized using:

$$w_{C,i}^{+25^o} \;=\; \widehat{w}_{C,i}^{+25^o} \,/\gamma \tag{14}$$

$$\vec{\mu}_{C,i}^{+25^o} \;=\; \vec{\mu}_{C,i}^{0^o} + \vec{\mu}_{\Delta,i}^{+25^o} \tag{15}$$

$$\left.\mathbf{\Sigma}_{C,i,(d,d)}^{+25^o}\right|_{d=1}^{D} \;=\; \left.\mathbf{\Sigma}_{C,i,(d,d)}^{0^o} \sigma_{\Delta,i,d}^{+25^o}\right|_{d=1}^{D} \tag{16}$$

where the non-diagonal elements of $\mathbf{\Sigma}_{C,i}^{+25^o}$ are set to zero and

$$\widehat{w}_{C,i}^{+25^o} \;=\; w_{C,i}^{0^o} \, w_{\Delta,i}^{+25^o} \tag{17}$$

$$\gamma \;=\; \sum_{i=1}^{N_G} \widehat{w}_{C,i}^{+25^o} \tag{18}$$

As can be seen, the γ is a scale factor used to ensure that synthesized weights sum to unity. We can of course use the above procedure to synthesize models for angles other than +25^o.

6.2. LinReg Technique

Let us suppose that we have the following multi-variate linear regression model:

$$\mathbf{Y} {=} \mathbf{X}\mathbf{B} \tag{19}$$

$$\begin{bmatrix} \vec{y}_1^{\,T} \\ \vec{y}_2^{\,T} \\ \vdots \\ \vec{y}_n^{\,T} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{\,T} & 1 \\ \vec{x}_2^{\,T} & 1 \\ \vdots & \\ \vec{x}_n^{\,T} & 1 \end{bmatrix} \begin{bmatrix} \beta_{(1,1)} & \cdots & \beta_{(1,D)} \\ \beta_{(2,1)} & \cdots & \beta_{(2,D)} \\ \vdots & \vdots & \vdots \\ \beta_{(D+1,1)} & \cdots & \beta_{(D+1,D)} \end{bmatrix} \tag{20}$$

where $n > D + 1$, with D being the dimensionality of each $\vec{y}$ and $\vec{x}$. $\mathbf{B}$ is a matrix of unknown regression parameters; under the sum-of-least-squares regression criterion, $\mathbf{B}$ can be found using [15]:

$$\mathbf{B} = \left(X^T X \right)^{-1} X^T Y \tag{21}$$

Given a set of *a priori* models (from group A), representing faces at 0^o and +25^o, we can thus find the relation between the means (and diagonal covariances) for the two angles; specifically, we find $\mathbf{B}_{\mu,i}$ and $\mathbf{B}_{\Sigma,i}$ (i=1,2,$\cdots$,N_G). We can then synthesize model parameters for +25^o [c.f. Eqn. (13)] from client C's 0^o model using:

$$\begin{aligned} w_{C,i}^{+25^o} &= w_{C,i}^{0^o} && (22) \\ \vec{\mu}_{C,i}^{+25^o} &= [\ (\vec{\mu}_{C,i}^{0^o})^T\ 1\ ]\ \mathbf{B}_{\mu,i} && (23) \\ \text{diag}(\mathbf{\Sigma}_{C,i}^{+25^o}) &= [\ \text{diag}(\mathbf{\Sigma}_{C,i}^{0^o})^T\ 1\ ]\ \mathbf{B}_{\Sigma,i} && (24) \end{aligned}$$

where the non-diagonal elements of $\mathbf{\Sigma}_{C,i}^{+25^o}$ are set to zero. It must be noted that unlike the UBMdiff technique (Section 6.1), there is no guarantee that the diagonal elements of $\mathbf{\Sigma}_{C,i}^{+25^o}$ are > 0; thus after synthesis, any diagonal elements which are ≤ 0 are set to a small positive value (1^{-25}). By the same token, the weights for the +25^o model are merely copied from the 0^o model (while this seems drastic, the weights have only a minor influence on performance [25]).

6.3. The Model Correspondence Problem

The UBMdiff and LinReg synthesis techniques pre-suppose that there is a correspondence between components of the client's 0^o model, the 0^o UBM, the +25^o UBM and all models for group A (loosely speaking, by correspondence we mean that corresponding components in all three models describe the same areas of the face). This is true when there is one Gaussian in each model (as for the PCA based system). However, under traditional training paradigms (as described in Section 5.1), this is generally not true when there is two or more Gaussians.

To address this issue, we propose the following modified training paradigm. Instead of training the +25^o UBM directly using the ML criterion, we instead adapt the 0^o UBM using a modified form of MAP estimation; moreover, whenever adapting any client model from any UBM, the modified MAP estimation is also used.

Traditional MAP estimation by itself will not help with the correspondence problem, as for GMMs it is a form of probabilistic clustering (albeit constrained clustering). During clustering, the Gaussians tend to "wander" around before converging to a solution[1]. We illustrate the wandering problem as follows: let's say we have a 32 Gaussian 0^o UBM and we adapt it to create a +25^o UBM; after convergence, it is quite possible for, say, the tenth Gaussian of the +25^o UBM to be the "closest" to the first Gaussian of the 0^o UBM; moreover, it is also possible to have more than one Gaussian in the +25^o UBM that is the "closest" to a given Gaussian in the 0^o UBM. Due to the "wandering" problem, there is no guarantee that the first Gaussian from the +25^o UBM corresponds to the first Gaussian from the 0^o UBM (or in other words, the first Gaussian from the +25^o UBM may be modeling a completely different area of the face when compared to the first Gaussian from the 0^o UBM).

Before describing the modification to the MAP estimation, let us first define a "parent UBM" as the UBM to be adapted and a "child UBM" as the UBM that resulted from adapting a "parent UBM"; in a similar vein, let us define a "parent Gaussian" as a Gaussian from the "parent UBM" and a "child Gaussian" as the Gaussian that resulted from a particular "parent Gaussian" through the process of adaptation. moreover, let us define the distance between two Gaussians as the Mahalanobis distance [9] between their means:

$$\mathcal{M}\left(\vec{\mu}_a,\ \vec{\mu}_b\right) = (\vec{\mu}_a - \vec{\mu}_b)^T \mathbf{\Sigma}_{all}^{-1} (\vec{\mu}_a - \vec{\mu}_b) \tag{25}$$

where $\mathbf{\Sigma}_{all}$ is the overall covariance matrix of the "parent UBM"; we shall assume that it is a diagonal matrix. It can be shown that the d-th diagonal element ($\mathbf{\Sigma}_{all,(d,d)}$) is found using:

$$\mathbf{\Sigma}_{all,(d,d)} = -\mu_{all,(d)}^2 + \sum_{i=1}^{N_G} w_i \left(\mathbf{\Sigma}_{i,(d,d)} + \mu_{i,(d)} \right) \tag{26}$$

where $\mu_{all,(d)}$ is the d-th element of $\vec{\mu}_{all}$, which is in turn found using $\vec{\mu}_{all} = \sum_{i=1}^{N_G} w_i \vec{\mu}_i$. Here, $\{w_i, \vec{\mu}_i, \mathbf{\Sigma}_i\}_{i=1}^{N_G}$ are the components of the "parent UBM".

Lastly, let us define a measure which will be used for checking whether any "child Gaussian" is closer to someone else's parent rather than its own parent:

$$\begin{aligned} \psi \ = \ & \sum_{i=1}^{N_G} \sum_{j=1}^{N_G} \mathcal{S} \left(k\mathcal{M}(\vec{\mu}_i^{\,child}, \vec{\mu}_i^{\,parent}) - \mathcal{M}(\vec{\mu}_i^{\,child}, \vec{\mu}_j^{\,parent}) \right) \\ & -2N_G \end{aligned} \tag{27}$$

where $k > 1$ and

$$\mathcal{S}(a) = \begin{cases} +1 & \text{if} \quad a > 0 \\ -1 & \text{if} \quad a \leq 0 \end{cases} \tag{28}$$

k designates how close a "child Gaussian" can be to someone else's parent; if k=2, then it is closer than two times the distance between the parent in question and the parent's true child.

To address the "wandering" problem we modify the EM algorithm for MAP estimation (shown in Appendix B) by introducing an early stopping criterion: from the second iteration onwards, we check if $\psi \neq -N_G^2$ after each maximization step; if the condition is satisfied we restore the parameters from the last iteration and deem that we have converged. The check is enabled from the second iteration onwards since we wish at least for some adaptation to occur (otherwise it would be possible for the "child UBM" to be the same as the "parent UBM"). In this work we use k=2 (choice based on preliminary experiments).

7. Augmenting Frontal Models

A composite model for client C is created by augmenting the client's frontal model ($\lambda_C^{0^o}$) as follows:

$$\begin{aligned} \lambda_C^{aug} &= \lambda_C^{0^o} \sqcup \lambda_C^{+60^o} \sqcup \lambda_C^{+40^o} \cdots \sqcup \lambda_C^{-40^o} \sqcup \lambda_C^{-60^o} \\ &= \sqcup_{i \in A} \lambda_C^i \end{aligned} \tag{29}$$

where

$$A = \{\ 0^o, +60^o, +40^o, +25^o, +15^o, -15^o, -25^o, -40^o, -60^o\ \} \tag{30}$$

and $\sqcup$ is an operator for joining GMM parameter sets. Let us suppose we have two GMM parameter sets, λ_x and λ_y, comprised of parameters for $N_{x,G}$ and $N_{y,G}$ Gaussians, respectively. The $\sqcup$ operator is defined as follows:

$$\begin{aligned} \lambda_z &= \lambda_x \sqcup \lambda_y \\ &= \{\alpha w_{x,i},\ \vec{\mu}_{x,i},\ \mathbf{\Sigma}_{x,i}\}_{i=1}^{N_{x,G}} \ \cup \ \{\beta w_{y,i},\ \vec{\mu}_{y,i},\ \mathbf{\Sigma}_{y,i}\}_{i=1}^{N_{y,G}} \end{aligned} \tag{31}$$

where:

$$\begin{aligned} \alpha &= N_{x,G}/(N_{x,G} + N_{y,G}) && (32) \\ \beta &= 1 - \alpha && (33) \end{aligned}$$

Here the non-frontal models can be synthesized from the client's frontal model using the UBMdiff or LinReg techniques (Section 6).

[1] It must be noted that this observed behaviour is counter-intuitive; it is under further investigation.

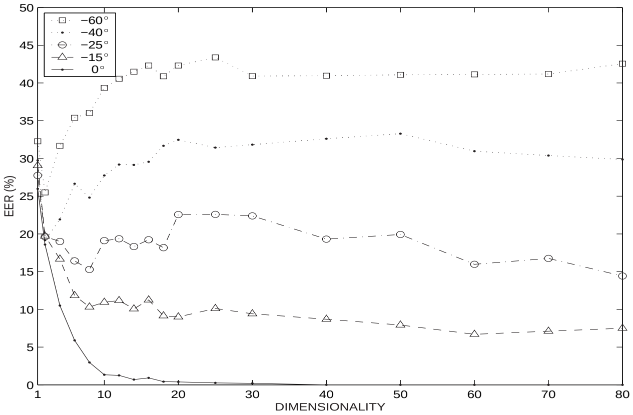


Fig. 3. Performance of PCA based system (trained on frontal faces) for increasing dimensionality and the following angles: -60^o, -40^o, -25^o, -15^o and 0^o (frontal).

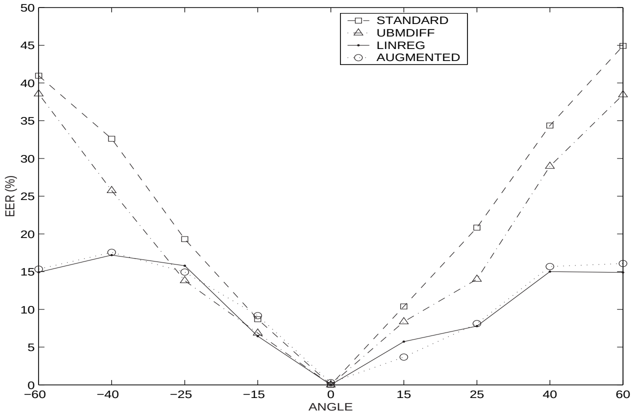


Fig. 4. Performance of various PCA based systems: standard, UBMdiff, LinReg and augmented; the standard system used original frontal client models only; UBMdiff and LinReg systems used client models synthesized specifically for a given test angle; the augmented system used client models comprised of original frontal and synthesized side models (via LinReg technique).

8. Experiments and Discussion

8.1. PCA Based System

In the first experiment we studied how the dimensionality of the feature vectors used in the PCA system affects robustness to varying pose. The higher the dimensionality, the more accurately the face image is represented; we conjecture that as more accurately the face is represented, the more the system will be affected by varying pose. Client models were trained on frontal faces and tested on faces from -60^o to +60^o views; impostor faces matched the testing view. Results for -60^o to 0^o are shown in Fig. 3 (the results for 0^o to +60^o, not shown here, have very similar trends).

As can be observed, a dimensionality of 40 is required to achieve perfect verification on frontal faces (this agrees with results presented in [26]). For non-frontal faces at $\pm 60^o$ and $\pm 40^o$, the error rate generally increases as the dimensionality is increased, saturating when the dimensionality is about 15; hence there is somewhat of a trade-off between the error rate on frontal faces and non-frontal faces, controlled by the dimensionality. Since in this work we are pursuing extensions to standard 2D approaches, the dimensionality has been fixed at 40 for further experiments (using a lower dimensionality of, say, 4, offers better (but still poor) performance for non-frontal faces, however it comes at the a cost of an EER of about 10% on frontal faces, which is unacceptable in real life applications).

In the second experiment we evaluated the performance of models synthesized using UBMdiff and LinReg techniques; The client models were synthesized for a given test angle; this pre-supposes that we know what the test angle is *a priori*, but is nevertheless useful for comparing performance with augmented models. As can be seen from the results presented in Fig. 4, both techniques perform better than the standard system and the LinReg technique offers significantly better performance than UBMdiff. We conjecture the reason for the betterness of the LinReg technique as follows: the UBMdiff technique only utilizes the difference between two general models, while the LinReg technique utilizes the differences between two sets of models (90 models for a frontal view and 90 models for a non-frontal view); in effect, the LinReg technique utilizes more information than the UBMdiff technique (in the form of 180 mean vectors instead of two) and is thus able to synthesize the non-frontal models more accurately. While the LinReg technique does not guarantee that valid covariance matrices will be generated, for the case of the PCA based system no such problem occurred; we conjecture that this is due to the constrained training strategy (Section 5.2), where client models inherited their covariance matrix from the UBM; in effect the LinReg technique uses information from two covariance matrices instead of 180.

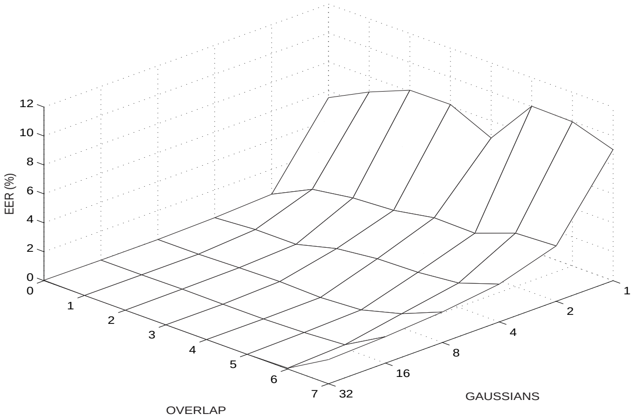


Fig. 5. Performance of standard DCTmod2 based system trained and tested on frontal faces, for varying degrees of overlap and number of Gaussians.

In the third experiment we augmented each client's frontal model with models (for the eight non-frontal views) synthesized by the LinReg technique; since each frontal model was constrained to have one Gaussian, each resulting augmented model had nine Gaussians. From the results shown in Fig. 4, we can see that there is little difference between using client models specifically synthesized for a given test angle and the augmented models, which cover all the test angles. These results thus support the use of frontal models augmented with synthesized models.

8.2. DCTmod2 Based System

In the first experiment we studied how the overlap setting in the DCTmod2 feature extractor and number of Gaussians in the classifier affects performance & robustness. Client models were trained on frontal faces and tested on faces at 0^o and +40^o views; impostor faces matched the testing view. Results are shown in Figs. 5 and 6.

As we can see, when testing with frontal faces, the general trend is that as the overlap increases more Gaussians are needed to decrease

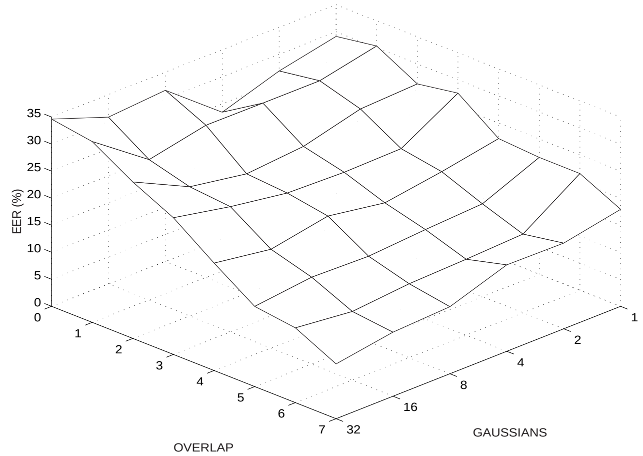


Fig. 6. Performance of standard DCTmod2 based system trained on frontal faces and tested on +40^o faces, for varying degrees of overlap and number of Gaussians.

the error rate (which can be interpreted as follows: the smaller the spatial area used by the features, the more Gaussians are required to adequately model the face). When testing with non-frontal faces, the general trend is that as the overlap increases, the lower the error rate; there is also a less defined trend when the overlap is 4 pixels or greater: the more Gaussians, the lower the error rate[2]. While not shown here, the DCTmod2 based system obtained similar trends for non-frontal views other than +40^o.

The best performance for +40^o faces is achieved with an overlap of 7 pixels and 32 Gaussians, resulting in an EER close to 10%. This is quite impressive. considering that the EER of the standard PCA based system is around 35%; for the PCA system utilizing synthesized models the EER is around 15%. The robustness of the standard DCTmod2/GMM system can be attributed to two aspects:

1. The small spatial area (especially with an overlap of 7) used by each feature vector, results in out-of-plane rotations having a smaller effect on DCTmod2 features when compared to PCA based features (which describe the entire face).
2. The loss of spatial relation between face characteristics (due to use of the GMM classifier), resulting in the "movement" of facial characteristics (due to out-of-plane rotations) having relatively little effect.

For further experiments we have chosen the configuration of 7 pixel overlap and 32 Gaussians. While this does not achieve perfect verification rate on frontal faces, the EER is quite low at 1.67%; moreover, as will be shown in the next experiment, the EER is close to zero when the modified MAP estimation is used (described in Section 6.3).

In the second experiment we evaluated the effects of modified MAP estimation. From the results presented in Fig. 7 we can see that utilizing the modified training has no adverse effects on the performance when compared to original MAP estimation.

In the third experiment we evaluated the performance of models synthesized via the UBMdiff technique, using both original and modified training. In order to provide a fair comparison with the LinReg technique in later experiments, synthesis of weights was not done; instead, the weights for non-frontal models were copied from the frontal model. As shown in Fig. 7, using original training causes the UBMdiff technique to fall apart (the results are worse than the standard approach); in contrast, using the UBMdiff technique with modified MAP estimation reduces the error rate in almost all cases. These results thus suggest that the model correspondence problem (described in Section 6.3) is effectively addressed via the modified MAP estimation; the results also suggest that the UBMdiff technique is useful for synthesizing models.

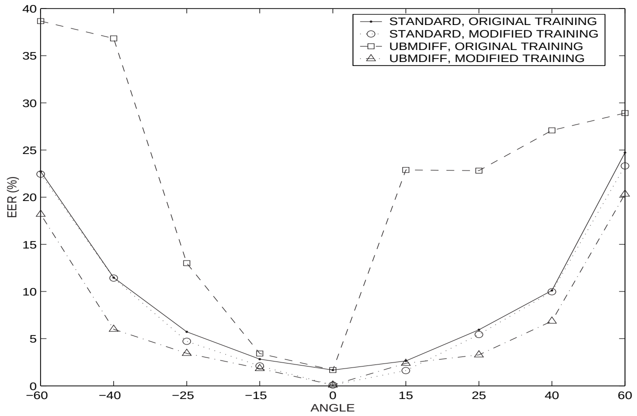


Fig. 7. Performance of various DCTmod2 based systems: standard (using original & modified training) and UBMdiff (also using original & modified training).

In the fourth experiment we evaluated the use of the LinReg technique for synthesizing models; results are presented in Fig. 8. It can be seen that the performance is worse than the UBMdiff technique; a possible cause of this has been alluded in Section 6.2: there is no guarantee that valid covariance matrices will be generated. Indeed, during model synthesis it was found that many elements of the covariance matrices had negative values, and were thus set to a small positive value; this obviously has the effect of making any model less precise, leading to worse performance.

In the fifth experiment we augmented each client's frontal model with models synthesized by the UBMdiff technique for the following angles: $\pm 60^o$, $\pm 40^o$ and $\pm 25^o$. Synthesized models for $\pm 15^o$ were not used since they provided no performance benefit over the 0^o model. Since each frontal model was set to have 32 Gaussians, each resulting augmented model had 224 Gaussians. From the results shown in Fig. 8, we can see that there is little difference between us-

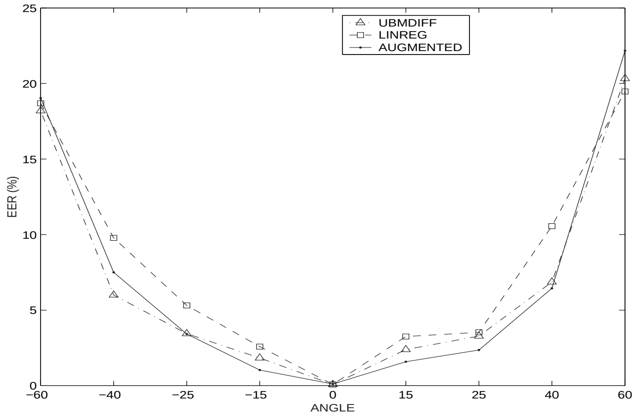


Fig. 8. Performance of various DCTmod2 based systems: UBMdiff, LinReg and augmented; UBMdiff and LinReg systems used client models synthesized specifically for a given test angle; the augmented system used client models comprised of original frontal and synthesized side models (via UBMdiff technique).

[2] This is true up to a point: eventually the error rate will go up as there will be too many Gaussians to train adequately with the limited amount of data.

ing client models specifically synthesized for a given test angle and the augmented models, which cover all the test angles. Like in the case for the PCA based system, these results support the use of frontal models augmented with synthesized models.

8.3. PCA/GMM vs DCTmod2/GMM

Since in this work we have evaluated two significantly different face verification systems (PCA based and DCTmod2 based), it would be interesting to compare their performance. The results shown in Fig. 9 (created by reusing results from previous experiments) suggest the following:

1. The standard DCTmod2/GMM system (trained on frontal faces) is less affected than the corresponding PCA/GMM system.
2. In almost all cases, frontal model augmentation has beneficial effects for both systems.
3. Except for the extreme views at $\pm 60^o$, the DCTmod2/GMM system using augmented models is more robust than the corresponding PCA/GMM system.

9. Conclusions and Future Work

In this work we proposed to address the problem of non-frontal face verification when only a frontal training image is available (e.g. a passport photograph) by augmenting a client's frontal face model with artificially synthesized models for non-frontal views. In the framework of a GMM based classifier, two techniques were proposed for the synthesis: UBMdiff and LinReg. Both techniques rely on *a priori* information and learn how face models for the frontal view are related to face models at a non-frontal view. The synthesis and augmentation approach was evaluated by applying it to two face verification systems: PCA based and DCTmod2 based; the two systems are a representation of holistic and non-holistic approaches, respectively.

Experimental results suggest that in almost all cases, frontal model augmentation has beneficial effects for both systems; they also suggest that the LinReg technique (which is based on multivariate regression of classifier parameters) is more suited to the PCA based system and that the UBMdiff technique (which is based on differences between two general face models) is more suited to the DCTmod2 based system. The results also support the view that the standard DCTmod2/GMM system (trained on frontal faces) is less affected by out-of-plane rotations than the corresponding PCA/GMM system; moreover, except for the extreme views at $\pm 60^o$, the DCTmod2/GMM system using augmented models is more robust than the corresponding PCA/GMM system.

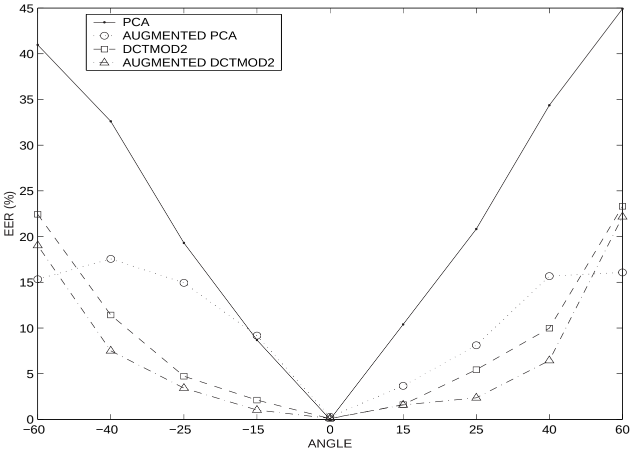


Fig. 9. Performance comparison of standard PCA, augmented PCA, standard DCTmod2 and augmented DCTmod2.

Currently in the DCTmod2/GMM approach each Gaussian often models disjoint face areas that are similar in texture (see Appendix A in [27]). This may not be optimal when dealing with out-of-plane face rotations as different parts of face may very well undergo different transformations. Better performance may be obtained if the Gaussians are constrained to model non-disjoint areas; to some extent this could be achieved by incorporating positional information in each feature vector (i.e. augmenting each DCTmod2 vector with the row and column of where it comes from); another possibility it to use a 2D Hidden Markov Model (HMM) based classifier [10, 26] in place of the GMM classifier.

Finally we note that, in the context of generative models (such as the GMM), there are probably more principled ways (than UBMdiff and LinReg) of utilizing *a priori* information; however, the techniques presented here show that it's possible to effectively utilize *a priori* information directly in the model domain, rather than in the image domain.

Appendix A. EM: Maximum Likelihood

Given a set of training vectors, $X = \{\vec{x}_i\}_{i=1}^{N_V}$, the GMM parameters ($\lambda$) are estimated using the Maximum Likelihood (ML) principle:

$$\lambda = \arg\max_{\hat{\lambda}} \; p(X|\hat{\lambda}) \tag{34}$$

The estimation problem can be solved using a form of the Expectation Maximization (EM) algorithm [6, 9]. The EM algorithm for GMMs is comprised of iterating two steps: the *expectation* step, followed by the *maximization* step. GMM parameters generated by the previous iteration (λ^{old}) are used by the current iteration to generate a new set of parameters (λ^{new}), such that:

$$p(X|\lambda^{new}) \geq p(X|\lambda^{old}) \tag{35}$$

The process is usually repeated until convergence (the parameters have not changed from one iteration to the next), or until the increase in the likelihood after each iteration falls below a pre-defined threshold, or until the number of iterations is equal to a pre-defined maximum. Reynolds [24] showed that the EM algorithm generally converges in 10 to 15 iterations, with further iterations resulting in only minor increases of the likelihood $p(X|\lambda)$; this has also been the authors' experience with various types of data. In our implementation we have therefore limited the number of iterations to 20. The algorithm is summarized as follows:

Expectation step:

for $k = 1, \cdots, N_G$: for $i = 1, \cdots, N_V$:

$$l_{k,i} = \frac{w_k \mathcal{N}(\vec{x}_i; \vec{\mu}_k, \mathbf{\Sigma}_k)}{\sum_{n=1}^{N_G} w_n \mathcal{N}(\vec{x}_i; \vec{\mu}_n, \mathbf{\Sigma}_n)} \tag{36}$$

for $k = 1, \cdots, N_G$:

$$L_k = \sum\nolimits_{i=1}^{N_V} l_{k,i} \tag{37}$$

$$\hat{w}_k = L_k / N_V \tag{38}$$

$$\hat{\vec{\mu}}_k = \frac{1}{L_k} \sum\nolimits_{i=1}^{N_V} \vec{x}_i \, l_{k,i} \tag{39}$$

$$\hat{\mathbf{\Sigma}}_k = \frac{1}{L_k} \left(\sum\nolimits_{i=1}^{N_V} \vec{x}_i \vec{x}_i^T l_{k,i} \right) - \hat{\vec{\mu}}_k \hat{\vec{\mu}}_k^T \tag{40}$$

Maximization step:

$$\{w_k, \vec{\mu}_k, \mathbf{\Sigma}_k\}_{k=1}^{N_G} = \{\hat{w}_k, \hat{\vec{\mu}}_k, \hat{\mathbf{\Sigma}}_k\}_{k=1}^{N_G} \tag{41}$$

The initial estimate (i.e. the seed) is typically provided by the k-means clustering algorithm [9]. It must be noted that the above implementation of EM can also be interpreted as an unsupervised probabilistic clustering procedure, with N_G being the assumed number of clusters.

Appendix B. EM: MAP Estimation

The main difference between ML and MAP estimation is in the use of *a priori* distribution ($f(\hat{\lambda})$) of the parameters to be estimated [c.f. Eqn. (34)]:

$$\lambda = \arg\max_{\hat{\lambda}} \; p(X|\hat{\lambda}) \; f(\hat{\lambda}) \tag{42}$$

The above estimation problem can be also solved using the EM algorithm, albeit in a different form to the one described in Appendix A; this form is often referred to as maximum *a posteriori* estimation [12, 25], and is summarized as follows.

Given UBM parameters $\lambda_{ubm} = \{\tilde{w}_k, \tilde{\vec{\mu}}_k, \tilde{\mathbf{\Sigma}}_k\}_{k=1}^{N_G}$ and a set of training feature vectors for a specific client, $X = \{\vec{x}_i\}_{i=1}^{N_V}$, the estimated weights ($\hat{w}_k$), means ($\hat{\vec{\mu}}_k$), and covariances ($\hat{\mathbf{\Sigma}}_k$) are found as per Eqns. (38)-(40). The maximization step (for $k = 1, \cdots, N_G$) is then defined as:

$$w_k = [\alpha\hat{w}_k + (1-\alpha)\tilde{w}_k]\,\gamma \tag{43}$$

$$\vec{\mu}_k = \alpha\hat{\vec{\mu}}_k + (1-\alpha)\tilde{\vec{\mu}}_k \tag{44}$$

$$\mathbf{\Sigma}_k = \left[\alpha\Big(\hat{\mathbf{\Sigma}}_k + \hat{\vec{\mu}}_k\hat{\vec{\mu}}_k^T\Big) + (1{-}\alpha)\Big(\tilde{\mathbf{\Sigma}}_k + \tilde{\vec{\mu}}_k\tilde{\vec{\mu}}_k^T)\Big)\right] {-} \vec{\mu}_k\vec{\mu}_k^T \tag{45}$$

where γ is a scale factor to make sure the weights sum to one. $\alpha{=}\frac{L_k}{L_k+r}$ is a data-dependent adaptation coefficient [L_k is found using Eqn. (37)], where r is a fixed relevance factor [25]; in our experiments we used r=256 (choice based on preliminary experiments).

As can be seen, the new parameters are simply a weighted sum of *a priori* statistics and new statistics. Here, α can be interpreted as the amount of faith we have in the new statistics. The choice of $\alpha{=}\frac{L_k}{L_k+r}$ causes the adaptation of only the Gaussians for which there is "sufficient" data; in other words, the MAP estimation approach for finding GMM parameters should be robust to limited amount of training data.

Since the ML EM algorithm for GMMs is a form of unsupervised probabilistic clustering, the MAP EM algorithm is also a form of unsupervised probabilistic clustering, albeit it is constrained.

References

[1] J. J. Atick, P. A. Griffin and A. N. Redlich, "Statistical Approach to Shape from Shading: Reconstruction of Three-Dimensional Face Surfaces from Single Two-Dimensional Images", *Neural Computation*, Vol. 8, 1996, pp. 1321-1340.

[2] D. Beymer and T. Poggio, "Face Recognition From One Example View", *Proc. 5th Int. Conf. Computer Vision (ICCV)*, Cambridge, 1995, pp. 500-507.

[3] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 15, No. 10, 1993, pp. 1042-1052.

[4] F. Cardinaux, C. Sanderson and S. Marcel, "Comparison of MLP and GMM Classifiers for Face Verification on XM2VTS", *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, 2003, pp. 911-920.

[5] L-F. Chen, H-Y. Liao, J-C. Lin and C-C. Han, "Why recognition in a statistics-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof", *Pattern Recognition*, Vol. 34, No. 7, 2001, pp. 1393-1403.

[6] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statistical Soc., Ser. B*, Vol. 39, No. 1, 1977, pp. 1-38.

[7] G. R. Doddington, M. A. Przybycki, A. F. Martin and D. A. Reynolds, "The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective", *Speech Communication*, Vol. 31, No. 2-3, 2000, pp. 225-254.

[8] B. Duc, S. Fischer and J. Bigün, "Face Authentication with Gabor Information on Deformable Graphs", *IEEE Trans. Image Processing*, Vol. 8, No. 4, 1999, pp. 504-516.

[9] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, John Wiley & Sons, USA, 2001.

[10] S. Eickeler, S. Müller and G. Rigoll, "Recognition of JPEG Compressed Face Images Based on Statistical Methods", *Image and Vision Computing*, Vol. 18, No. 4, 2000, pp. 279-287.

[11] S. Furui, "Recent Advances in Speaker Recognition", *Pattern Recognition Letters*, Vol. 18, No. 9, 1997, pp. 859-872.

[12] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, 1994, pp. 291-298.

[13] R. C. Gonzales and R. E. Woods, *Digital Image Processing*, Addison-Wesley, Reading, Massachusetts, 1993.

[14] R. Gross, J. Yang and A. Waibel, "Growing Gaussian Mixture Models for Pose Invariant Face Recognition", *Proc. 15th Int. Conf. Pattern Recognition*, Barcelona, 2000, pp. 1088-1091 (Vol. 1).

[15] D.-Y. Huang and K.-C. Liu, "Some variable selection procedures in multivariate linear regression models", *J. Statistical Planning and Inference*, Vol. 41, 1994, pp. 205-214.

[16] M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R. P. Würtz and W. Konen, "Distortion Invariant Object Recognition in the Dynamic Link Architecture", *IEEE Trans. Computers*, Vol. 42, No. 3, 1993, pp. 300-311.

[17] M. W. Lee and S. Ranganath, "Pose-invariant face recognition using a 3D deformable model", *Pattern Recognition*, Vol. 36, No. 8, 2003, pp. 1835-1846.

[18] T. Maurer and C. v.d. Malsburg, "Learning Feature Transformations to Recognize Faces Rotated in Depth", *Proc. Int. Conf. Artificial Neural Networks (ICANN)*, Paris, 1995, pp. 353-358.

[19] K. Messer et al., "Face Verification Competition on the XM2VTS Database", *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, 2003, pp. 964-974.

[20] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, 1997, pp. 696-710.

[21] P. Niyogi, F. Girosi and T. Poggio, "Incorporating Prior Information in Machine Learning by Creating Virtual Examples", *Proceedings of the IEEE*, Vol. 86, No. 11, 1998, pp. 2196-2209.

[22] A. Pentland, B. Moghaddam and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition", *Proc. Int. Conf. Computer Vision and Pattern Recognition*, Seattle, 1994, pp. 84-91.

[23] P. J. Phillips, H. Moon, S. A. Rizvi and P. J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, 2000, pp. 1090-1104.

[24] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", *Technical Report 967*, Lincoln Laboratory, Massachusetts Institute of Technology, 1993.

[25] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, Vol. 10, No. 1-3, 2000, pp. 19-41.

[26] F. Samaria, *Face Recognition Using Hidden Markov Models*, PhD Thesis, University of Cambridge, 1994.

[27] C. Sanderson, "Face Processing & Frontal Face Verification", IDIAP-RR 03-20, Martigny, Switzerland, 2003. (see *www.idiap.ch*)

[28] C. Sanderson and S. Bengio, "Robust Features for Frontal Face Authentication in Difficult Image Conditions", *Proc. 4th Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, Guildford, 2003, pp. 495-504.

[29] C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes", *Pattern Recognition Letters*, Vol. 24, No. 14, 2003, pp. 2409-2419.

[30] M. Turk and A. Pentland, "Eigenfaces for Recognition", *J. Cognitive Neuroscience*, Vol. 3, No. 1, 1991, pp. 71-86.

Distortion -Tolerant Iris Recognition Using Advanced Correlation Filters

B.V.K. Vijaya Kumar and Jason Thornton
Dept. of ECE, Carnegie Mellon University, Pittsburgh, PA 15213
kumar@ece.cmu.edu, jthornto@andrew.cmu.edu

Abstract

The iris is potentially a very distinct and useful biometric because of its intricate patterns. One way to extract information about these patterns is through texture analysis using Gabor wavelets. However, such analysis does not explicitly handle within-class variation among iris textures. Correlation filters offer a different approach by working on the spatial frequency spectrum of an iris. Correlation filters can be designed to give sharp peaks in response to authentic iris images and no such peaks in response to impostor iris images. The spatial frequencies that make up a correlation filter can be optimally selected to maintain these peaks in the presence of within-class distortions. This paper compares the iris recognition capability of the Gabor wavelet analysis method and the correlation filter method on an iris image set with a range of introduced distortions.

1. Introduction

Biometrics are useful in distinguishing between subjects for identity recognition purposes. A good biometric should be present throughout the lifetime of an individual, be distinct enough to identify one individual from others with certainty, and be readily accessible to some kind of outside sensor. The iris has these properties, and therefore has potential as a very effective biometric.

The iris is the colored part of the eye that surrounds the pupil and dilates or constricts the pupil opening (see Fig. 1). The visual patterns of the iris are set before birth, and empirical evidence suggests that they remain stable over a person's lifetime. The patterns are thought to be unique to each eye (right and left) of every individual, providing enough information to recognize someone with confidence. In addition, an iris image can be recorded externally and without contact with a subject (in contrast to retinal scanning), although subject cooperation is generally necessary to get an image of good enough quality.

An iris recognition system must take an eye image, separate the iris from the rest of the image, and extract information from the iris that can be used to identify it. Daugman has pioneered the use of Gabor wavelet analysis to characterize the local textures of an iris [1].

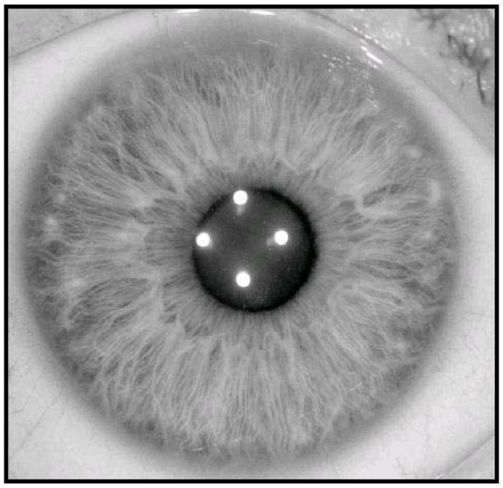


Figure 1. Example of iris image

This method involves filtering the iris image with 2-D Gabor wavelets at different scales, orientations, and locations. Only the phase values of the wavelet processing are kept as reliable information, and are quantized into two bits per phase value. These bits taken together form the "iris code", and can be stored as the representation of an iris. Because an iris code is created from a series of local texture analysis computations, it is not completely corrupted when part of the iris is occluded or otherwise invalid. In this sense, the iris code degrades gracefully with partial occlusion (like from an eyelid). However, this technique does assume that local textures remain highly consistent across within-class images, which can be problematic in the case of within-class distortions.

The use of correlation filters [2] may offer an attractive alternative for the task of distortion-tolerant iris recognition. Correlation filters are designed in the spatial frequency domain, one for each class. When an input image belonging to the authentic class is filtered, a peak results in the correlation output; when the input image does not belong to the authentic class, filtering should produce no such peak. Correlation filters offer several advantages. First, they are naturally shift-invariant, so translation of the input image does not affect recognition.

Thus, the input images do not have to be centered. Also, they can be designed, using multiple within-class images, to handle within-class distortion. In addition, the performance of correlation filters degrades gracefully in the presence of noise and occlusions. This is because correlation is an integrative operation and thus no particular input image pixels are important by themselves. The authors have recently reported results from preliminary studies on iris recognition using correlation filters [13].

This paper focuses on the use of correlation filters (specifically the type of correlation filters that allow us to optimally trade-off correlation peak sharpness for noise tolerance) as an alternative to Gabor wavelet analysis in performing iris recognition. Section 2 discusses the pre-processing employed. Section 3 explains our implementation of the Gabor wavelet analysis technique to produce the iris codes introduced by Daugman. Section 4 discusses distortion-invariant correlation filters and their design. Section 5 presents our testing procedure and the results of both recognition algorithms; this includes verification (also known as 1:1 matching where the system compares the live biometric to a stored one and accepts or rejects the claimed identity) and identification (also known as 1:*N* matching where the live biometric is compared to a database of biometrics to determine the identity of the subject). Section 6 provides a summary.

2. Iris image preprocessing

In order to use only the texture information from the iris patterns and to avoid unreliable information from uninteresting regions in the eye image (such as the pupil), the iris must be separated from the rest of the eye. It is important to segment all irises into a normalized form in order to make analysis consistent.

Segmentation and normalization are required preprocessing for the iris code method. On the other hand, correlation filters are shift-invariant and so are capable of operating on entire eye images without explicitly defining the iris region. However, the preprocessing described in this section does improve correlation filter performance for two reasons. If correlation filters are designed to recognize entire eye images, they may emphasize information that is not specific to the iris and therefore not a stable characteristic (such as the contour of the eyelid or the presence of eyelashes). In addition, the correlation filters applied in this paper are not designed to handle scale changes. The normalization of the iris assures that scale changes will not affect performance. For these reasons, all iris images used by either algorithm are preprocessed as follows.

2.1. Detecting iris boundaries

Both the inner and outer iris boundaries can be modeled as circles with a fair amount of accuracy. This simplifying assumption helps to make the segmentation process computationally manageable. However, the two circular boundaries need not be concentric. The location of the boundaries are indicated by a sudden change in image intensity from darker (inside the circular boundary) to lighter (outside the boundary), and can be found using some form of radial gradient operator. This approach towards locating the iris boundaries is very effective and has been described by Daugman in his earlier work [3].

However, finding the boundaries within a reasonable computation time is tricky. This is because for each boundary, the algorithm must find the center location that yields the highest radial gradient. This is not a trivial search. In order to find the maximum radial gradient, the algorithm must approximate circular integration of the image for every possible combination of center and radius values. The standard way to accomplish this is to perform a polar transform around every possible center, and project each transform onto its radial axis. But the interpolation involved with a discrete polar transform is computationally costly, even at coarse scales.

Instead, our algorithm relies on cross-correlation to approximate circular integration. Take an image that contains only a single circle of radius ρ against a zero-valued background, and compute its inner product with an iris image; the result can be considered an approximation to integrating the iris image along a circular contour with radius ρ. If the cross-correlation is computed, it gives a set of shifted inner products. So the result is an approximation of circular integration at radius ρ across all center locations. If the iris image is cross-correlated with a set of circles of every possible radius (as shown in Fig. 2), all the necessary circular integrations have been approximated.

Computing a series of cross-correlations is more efficient than computing a series of polar transform interpolations. This is because cross-correlation can be computed in the frequency domain via a simple conjugate multiplication, and the Fast Fourier Transform algorithm offers an efficient way to transform in to and out of the frequency domain.

Our algorithm first downsamples an iris image to 100 by 100 pixels. Then it computes the cross-correlation between the coarse image and a bank of 100 circles of different radii. A gradient operator is applied to the resulting circular integrations to find the maximum radial gradients, one for the inner boundary and one for the outer boundary.

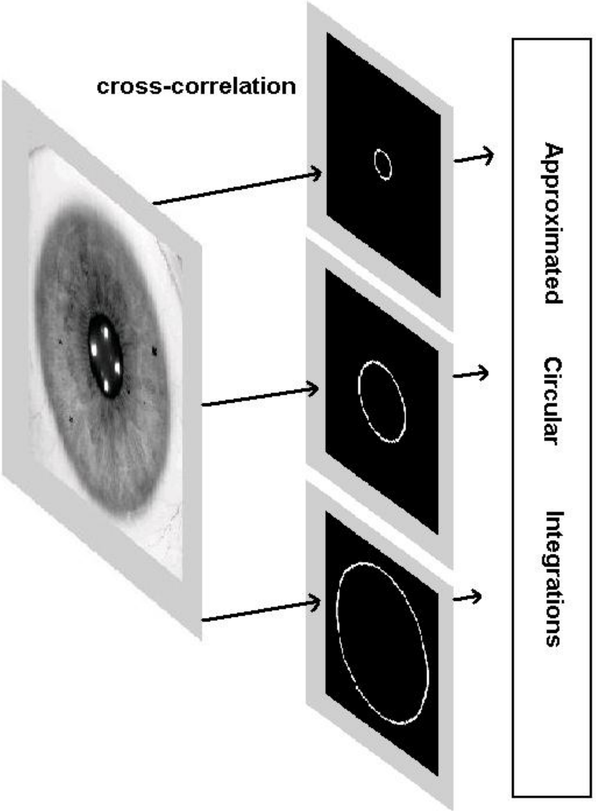


Figure 2. Computing circular integrations using cross-correlations

2.2. Normalized segmentation

When the iris boundaries have been selected, the final step is to segment the iris. The goal is to project the iris onto an unwrapped polar coordinate system with a uniform radius. At each discrete angle in the image, there is a radial line from the center of the pupil that crosses the inner and outer boundaries of the iris. The length of the line segment that intersects the iris varies with angle, because the two boundaries are not concentric. For radial normalization, the same number of intensity samples are taken along every radial line, equally distributed. The resulting unwrapped and segmented iris has a uniform radius across all angles, as depicted in Fig. 3.

This iris segmentation method automatically takes care of translation because the iris region is located by boundary detection. It also takes care of scale changes because the radius is normalized. All irises are projected onto the same rectangular area, which makes analysis and comparison between irises possible.

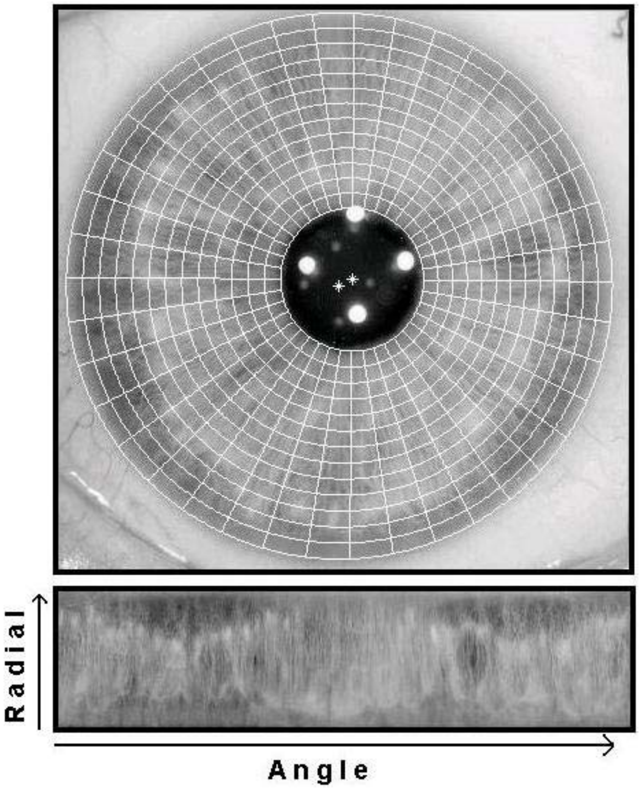


Figure 3. Example of iris segmentation
The top image displays the detected inner and outer iris boundaries, which have different centers. The grid covering the area between the boundaries shows some of the radial lines that are sampled (at a resolution that normalizes their length). The bottom image shows the resulting unwrapped iris, with uniform normalized radius.

3. Using Gabor wavelets to create iris codes

Daugman has popularized the use of Gabor wavelet decomposition to characterize the local textures of an iris [3]. "Iris codes" can be created by quantizing the results of this texture analysis. We implemented an algorithm that uses Gabor wavelet texture analysis to produce a version of iris codes, in order to compare this approach to our correlation filter approach. Our implementation is outlined below.

We created a complex-valued, 2-D Gabor wavelet for the polar coordinate system of the unwrapped irises defined over angle Φ and radius ρ:

$$\tilde{g}(\phi,\rho)=\exp\left(-\frac{1}{2}\left[\left(\frac{\phi}{\sigma_{\varphi}}\right)^2+\left(\frac{\rho}{\sigma_{\rho}}\right)^2\right]-j2\pi\omega\phi\right)$$

where ω controls modulation, and any constant amplitude terms are unimportant. This is considered the mother wavelet, and from it we generated a self-similar family of wavelets at 8 different orientations and 4 different scales. These 32 wavelets are placed at different locations in the segmented iris plane to compute the complex projection of local parts of the iris onto the wavelets, as given by:

$$\text{Proj} = \int_{\phi}\int_{\rho} I(\phi,\rho)\,\tilde{g}(\phi-\phi_0,\rho-\rho_0)\,\rho d\rho d\phi$$

where $I(\phi,\rho)$ represents the segmented iris plane, with ϕ_0 and ρ_0 determining the spatial location of the wavelet in that plane. During analysis, wavelets of different scales are placed at different locations across the iris plane, with smaller wavelets at dense distributions and larger wavelets at sparse distributions. In total, slightly over 1000 wavelet projections are calculated.

Each projection yields one complex number, and the phase of this complex number is the important part [3]. As suggested by Daugman, we quantize each projection's phase into two bits based on the complex-plane quadrant it occupies. The array of 2118 bits that results from quantized Gabor wavelet projections represent our version of the iris code. In this approach to iris recognition, the iris code is meant to be a complete and unique representation of each iris.

Checking for degree of match between two iris codes is simple. The metric used is based on the Hamming distance, the number of corresponding bits that differ between the codes. The match metric m is computed as

$$m = 1 - \frac{\text{Hamming Distance}}{\text{Total Number of Bits}}$$

and ranges from approximately 0.5 when the codes are statistically independent to 1 when the codes are a perfect match.

An iris code is generated from a single iris image. If multiple reference images of the same iris are available, an iris code can be generated for each reference image. When dealing with a set of training images, we create and store an iris code for every reference image. Then during iris recognition, all the codes stored for one iris class are compared against the input test iris code to check for the closest match.

4. Using correlation filters for iris recognition

Correlation filters are based on the concept that filtering images in their spatial frequency domain is an effective way to recognize specific patterns. Their use for biometric recognition has been previously explored by Kumar et al [4]. Correlation filters are applied as shown in Fig. 4. The input image is first decomposed into its spatial frequencies by a Discrete Fourier Transform. This is accomplished using the computationally efficient Fast Fourier Transform (FFT) algorithm. The transformed image is multiplied by the complex conjugate of the correlation filter (note that conjugate multiplication in this domain equates to cross-correlation in the original spatial domain). Then it is transformed back to the original domain through an Inverse Fast Fourier Transform (IFFT). The result is referred to as the correlation plane.

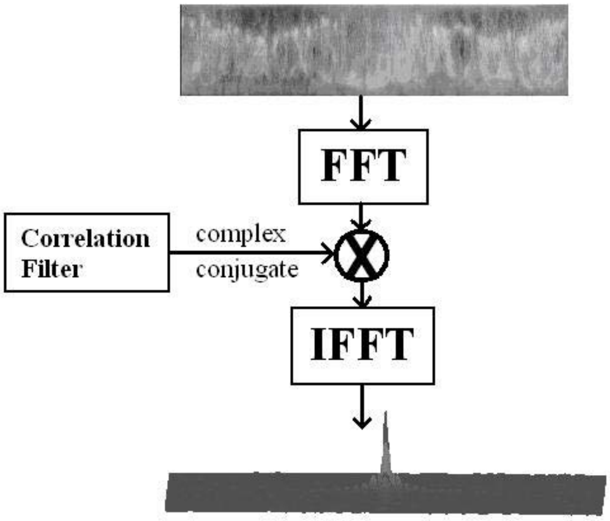


Figure 4. The application of a correlation filter to an iris image. The resulting correlation output produces a peak for authentic and no such peak for impostors.

If the input image is authentic (i.e., it contains the target pattern that the correlation filter was designed to recognize), the correlation plane should show a distinct correlation peak. An ideal correlation peak approximates a 2-D delta function. If the input image is an imposter (does not contain the target pattern), no distinct correlation peak should exist. The sharpness of the peak in the correlation plane is measured with the peak-to-sidelobe (PSR) ratio. The PSR metric takes several different forms, but for the purpose of this paper it is defined as

$$\text{PSR}=\frac{(\text{Peak-Mean})}{\text{Standard Deviation}}$$

High PSR in the correlation plane is considered an indication of match, while low PSR indicates no match. One natural advantage of correlation filters is their shift invariance. A shift in the input causes a corresponding shift in the correlation plane, but the peak values as well as mean and standard deviation values do not change and so the peak-to-sidelobe ratio remains the same.

4.1. Composite correlation filters

The simplest type of correlation filter is the Matched Filter [5], which is optimal for detecting an exact target pattern in additive white noise. However, this type of filter does not perform well in the presence of any kind of distortion. Casasent and Hester established the concept of composite correlation filters as a way to handle detection of multiple images belonging to the same class [6]. The goal is to design a filter that gives good recognition peaks for all reference images of the same class in a training set.

One effective version of a composite correlation filter is the Minimum Average Correlation Energy (MACE) filter [7]. The MACE filter is designed to give a specified peak value in the correlation plane for every reference image in the training set, while minimizing the average energy of the correlation plane. This minimization has a closed-form solution, making the filter design process straightforward. The result is that the recognition peaks are very sharp, giving high PSR scores for each of the reference images from the authentic class.

The MACE filter exhibits great discrimination performance for the reference images used to design it, but does not take into account within-class noise. However, the filter design can be adjusted to accommodate noisy versions of authentic images. Minimum Variance Synthetic Discriminant Function (MVSDF) filters are designed to perform optimally in the presence of noise [8]. Unfortunately, MVSDF filters do not offer good discrimination although they exhibit excellent distortion tolerance. MVSDF and MACE filters exhibit opposing attributes in that the former emphasizes low spatial frequencies whereas the latter emphasizes high spatial frequencies. Refregier showed that the two performance criteria, discrimination ability and noise tolerance, can be traded off optimally [9]. This led to the design of the Optimal Trade-off Synthetic Discriminant Function (OTSDF) filter [10], which we use in this application.

4.2. OTSDF filter design

Let α represent a trade-off parameter ranging from 0 to 1, determining the relative importance of noise tolerance to discrimination ability over the reference set. Then the design of the OTSDF filter is given by the closed-form solution:

$$\mathbf{f}=\left[\alpha\mathbf{D}+\sqrt{1-\alpha^2}\mathbf{C}\right]^{-1}\mathbf{m}$$

where **D** is a diagonal matrix containing the average power spectrum of the reference images along its diagonal, **m** is the mean of the DFTs of the reference images, in the shape of a column vector, and **C** is the power spectral density of the expected within-class noise. If it is assumed that the noise is white, as we do here, the formula simplifies to

$$\mathbf{f}=\left[\alpha\mathbf{D}+\sqrt{1-\alpha^2}\mathbf{I}\right]^{-1}\mathbf{m}$$

with **I** being the identity matrix. We set the trade-off parameter α to 0.5 when creating our filters for this application.

The OTSDF correlation filter is designed to recognize multiple images of the same iris, and to degrade gracefully in the presence of noise. This makes it well-suited to the task of recognizing an entire class of image patterns (in this case, iris patterns belonging to the same iris) while discriminating against imposter images. The filter does especially well if the reference images used to train it reflect an accurate range of within-class distortions. Also, an iris image with in-plane rotation shows up as a cyclical shift in the segmented iris. Correlation filters are shift-invariant, so they should be more capable of recognizing rotated irises.

5. Experimental results

The purpose of our numerical experiments was to test the two iris recognition methods across a range of within-class distortions. However, rigorous evaluation with iris images is difficult because there are no publicly available iris databases. For our tests, we created a data set of iris images by artificially introducing four different types of possible distortion. The iris segmentation process already takes care of scale changes and translation, so we selected distortions other than these that can make iris recognition difficult. They are listed below:

Rotation of eye: This can occur with tilting of the subject's head or the camera.

Partial occlusion by eyelid: This is modeled by placing a rectangle in the upper or lower portion of the eye image to obstruct part of the iris.
Random noise: Gaussian noise added to achieve a specific Signal-to-noise Ratio (SNR)
Nonlinear contortion: A slightly nonlinear stretching of the iris along the radial axis to simulate contortion caused by pupil movement

Fig. 5 shows examples of these individual distortions on one iris image. We used varying levels of each type of distortion to create separate images for the same iris. In addition, we used every pair-wise combination of distortions to create more images. The different distortion types and levels that were applied for every iris class are listed below:

Rotation only:	6 levels
Occlusion only:	8 levels
Gaussian noise only:	3 levels
Contortion only:	6 levels
Rotation and Occlusion:	30 combinations
Rotation and Gaussian noise:	10 combinations
Rotation and Contortion:	20 combinations
Occlusion and Gaussian noise:	12 combinations
Occlusion and Contortion:	24 combinations
Gaussian noise and Contortion:	8 combinations

So each iris class in our dataset consists of a total of 128 images (including the original), all distorted versions of the same iris. We created 45 iris classes for the dataset, based on high resolution iris images provided by Miles Research Lab [11].

Seven images from each class, representing the range of distortions, were set aside as training images. These images were used to create the iris codes and correlation filter used to characterize each iris class. This left 121 images in each class for testing purposes.

In testing, every image was compared to the iris codes and correlation filters stored for all 45 classes (considered an authentic to the class which it belonged, an imposter to all other classes). This generated a total of 5,445 authentic comparisons and a total of 239,580 imposter comparisons. Every comparison was scored by the respective match metrics of each algorithm (Hamming distance metric for iris codes, PSR for correlation filters).

5.1. Verification results

When iris recognition is used for verification, the subject claims an identity to be verified. The algorithm has to give a yes/no decision based on the match score, and therefore has to use some threshold value that separates authentics from imposters. This threshold value is universal and not specific to each class. We evaluated each algorithm in the context of verification by calculating the Equal Error Rate (EER). When a threshold value is selected that gives the same rate of false acceptances and false rejections, this error rate is the EER. We separated the test data based on type of distortion, and calculated the EER for each type. The results are shown in Table 1. Clearly, rotated irises were the largest challenge and affected performance the most. The other distortions had little or no impact on verification. Also, correlation filters performed much better than our version of iris code for rotation distortions in iris images.

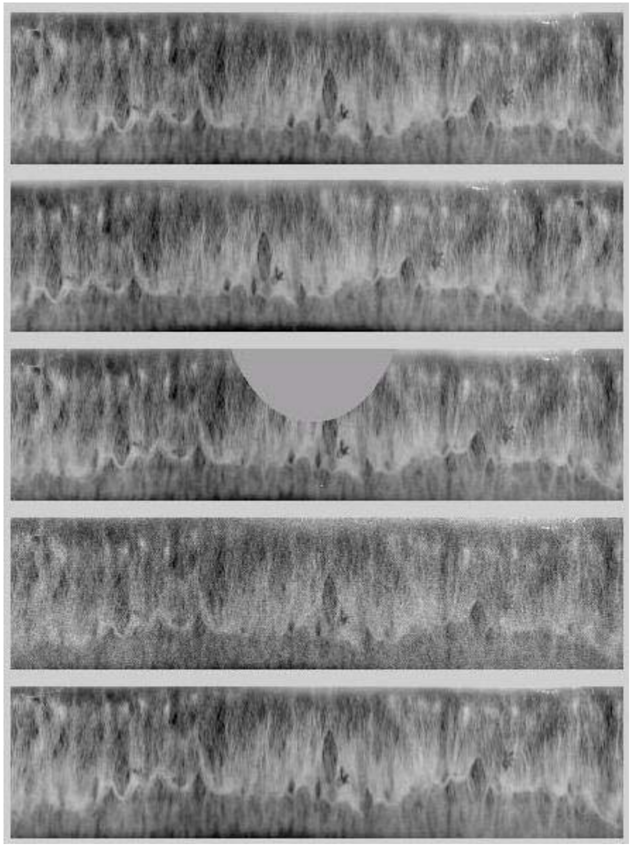


Figure 5. Examples of single distortions
From top to bottom: original (no distortion), rotation, partial occlusion, Gaussian noise, and nonlinear contortion outward along radial axis

5.2. Identification results

When iris recognition is used for identification, the algorithm has a different task. Instead of confirming the subject’s identity, it has to search for the identity among all stored classes. Identification is only successful if the match score returned for the authentic comparison is higher than the match scores returned for all other comparisons. We measured the identification ability of the algorithms with Cumulative Match Characteristic curves (CMC). This curve plots the ratio of test images that are correctly identified in top *k* match scores, as a function of *k*. If the identification algorithm performs well, a large ratio of test images will be correctly identified by the single greatest match score; this means the CMC curve will start very high and approach 1 quickly. As identification performance degrades, the area under the CMC curve decreases.

Table 1. Equal Error Rates

Distortion Type	Gabor Wavelet / Iris Code	Correlation Filter
Rotation (R)	41.3 %	8.9 %
Occlusion (O)	0 %	0 %
Noise (N)	0 %	0 %
Contortion (C)	0.06 %	0 %
(R) and (O)	41.2 %	9.1 %
(R) and (N)	42.4 %	14.6 %
(R) and (C)	43.4 %	9.7 %
(O) and (N)	0.002 %	0 %
(O) and (C)	0.13 %	0 %
(N) and (C)	0.83 %	0.52 %

Fig. 6 shows the CMC curves for both algorithms, calculated across all test images. The curve for correlation filters starts higher because 94.8% of test images are correctly identified by the top match score, as opposed to 52.6 % for iris codes based on Gabor wavelet analysis.

As with verification, it is the distorted images with rotation that prove difficult to identify. In fact, if all test images with some degree of rotation are disregarded (leaving only combinations of the other three distortion types), both algorithms give perfect identification with the top match score. So the identification results on non-rotated images are actually better than the verification results on those same images (because verification error is not always zero). This suggests that, excluding rotation, all classes have good separation between authentic and imposter match scores, although the threshold of separation varies between classes.

6. Conclusions

Both iris code and correlation filter algorithms give fairly good verification and identification performance on three of four distortion types, but rotation degrades performance significantly. The iris code method does not do well under rotation because it is based on local texture analysis. Rotating the iris causes local textures to shift to new spatial regions, rearranging the bits that make up an iris code. One possible way to overcome this problem is to store iris codes for rotated versions of the same iris image. In fact, one of the reference images from each class in our experiment was rotated, but it did not seem to help iris code performance much. The problem is that the reference image and the test image must be rotated by almost the same degree to get good recognition. The correlation filter showed better performance on rotated images because of its shift invariance. If the interpolation involved in segmenting the iris allowed for exact shifts among the discrete-valued iris patterns, the error rates of correlation filters on these images would approach zero.

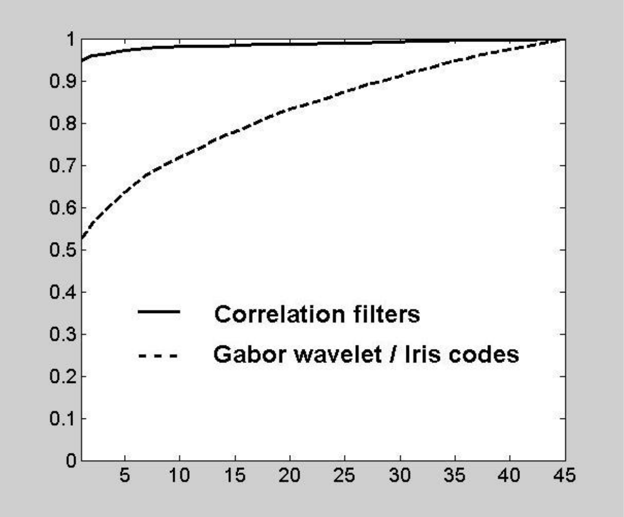


Figure 6. CMC curves for each algorithm
The horizontal axis is the number of top match scores that are considered, from 1 to all 45 scores. The vertical axis is the ratio of all test images which have their authentic score among the scores considered. Solid line shows the results for the correlation filters and dashed lines show the results for our implementation of the iris code method.

Overall, the OTSDF correlation filter demonstrated better verification and identification. This can be attributed to the design approach, which seeks good discrimination across the range of within-class distortion. It does not assume within-class consistency, as local texture analysis does. Instead, the reference images are used to determine which parts of the frequency spectrum are consistent enough to be useful for recognition. It remains to be seen if this approach scales effectively to a large, comprehensive database of iris images. We are in the process of creating our own database using iris camera equipment.

As a final note on algorithm design, the iris code method uses a level of phase quantization that is meant to simplify storage and computation for practical implementation. This paper does not focus on issues of practical implementation. But the effect of quantization on the design and use of correlation filters has been studied recently [12].

This research is supported in part by by the Army Research Office (ARO) through its support to the Center for Communications and Computer Security (C3S) at Carnegie Mellon University.

References

[1] J.G. Daugman, "High Confidence Visual Recognition of Persons by a Test of Statistical Independence," *IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 15, no. 11*, 1993, pp. 1148-1161.

[2] B.V.K. Vijaya Kumar, "Tutorial Survey of Composite Filter Designs for Optical Correlators," *Applied Opt., Vol. 31*, 1992, pp. 4773-4801.

[3] J.G Daugman, "Recognizing Iris Texture by Phase Demodulation," *IEEE Colloquium on on Image Processing for Biometric Measurement*, London, UK, 4-20-1994, pp. 2/1-2/8.

[4] B.V.K. Vijaya Kumar, M. Savvides, K. Venkataramani, C. Xie, "Spatial Frequency Domain Image Processing for Biometric Recognition," *Proc. Of Intl. Conf. on Image Processing (ICIP)*, Rochester, NY, Sept. 2002.

[5] A. Vanderlugt, "Signal Detection by Complex Spatial Filtering," *IEEE Trans. Inf. Theory 10*, 1964, pp. 139-145.

[6] C.F. Hester and D. Cassasent, "Multivariant Technique for Multiclass Pattern Recognition," *Appl. Opt., Vol. 19*, 1980, pp.1758-1761.

[7] A. Mahalanobis, B.V.K. Vijaya Kumar, D. Cassasent, "Minimum Average Correlation Energy Filters," *Appl. Opt., Vol. 26*, 1987, pp. 3630-3633.

[8] B.V.K. Vijaya Kumar, "Minimum Variance Synthetic Discriminant Functions," *Opt. Soc. Am. A, Vol. 3*, 1986, pp. 1579-1584.

[9] Ph. Refregier, "Optimal Trade-off Filters for Noise Robustness, Sharpness of the Correlation Peak, and Horner Efficiency," *Optics Letters, Vol. 16*, 1991, pp. 829-831.

[10] B.V.K. Vijaya Kumar, D.W. Carlson, and A. Mahalanobis, "Optimal Trade-off Synthetic Discriminant Function Filters for Arbitrary Devices," *Optics Letters, Vol. 19*, 1994, pp. 1556-1558.

[11] Miles Research Laboratory, San Diego, CA www.milesresearch.com.

[12] M. Savvides and B.V.K. Vijaya Kumar, "Quad Phase Minimum Average Correlation Energy Filters for Reduced Memory Illumination Tolerant Face Authentication," *Proc. 4th Int. Conf. On Audio and Video Based Biometric Person Identification (AVBPA 2003)*, Guildford, U.K., June 2003, pp. 19-26.

[13] B.V.K. Vijaya Kumar, C. Xie, J. Thornton, "Iris Verification Using Correlation Filters," *Proc. 4th Int. Conf. On Audio and Video Based Biometric Person Identification (AVBPA 2003)*, Guildford, U.K., June 2003, pp. 697-705.

Individual Recognition Using Gait Energy Image

Ju Han and Bir Bhanu
Center for Research in Intelligent Systems
University of California, Riverside, California 92521, USA
{jhan,bhanu}@cris.ucr.edu

Abstract

In this paper, we propose a new spatio-temporal gait representation, called Gait Energy Image (GEI), to characterize human walking properties for individual recognition by gait. To address the problem of the lack of training templates, we generate a series of new GEI templates by analyzing the human silhouette distortion under various conditions. Principal component analysis followed by multiple discriminant analysis are used for learning features from the expanded GEI training templates. Recognition is carried out based on the learned features. Experimental results show that the proposed GEI is an effective and efficient gait representation for individual recognition, and the proposed approach achieves highly competitive performance with respect to current gait recognition approaches.

1. Introduction

Current human recognition methods, such as fingerprints, face or iris biometrics, generally require a cooperative subject, views from certain aspects and physical contact or close proximity. These methods can not reliably recognize non-cooperating individuals at a distance in real-world changing environmental conditions. Moreover, in various applications of personal identification, many established biometrics can be obscured. Gait, which concerns recognizing individuals by the way they walk, has been an important biometric without the above-mentioned disadvantages.

In this paper, we propose a new spatio-temporal gait representation, Gait Energy Image (GEI), for individual recognition. Unlike other gait representations [8, 4] which consider gait as a sequence of templates (poses), GEI represents human motion sequence in a single image while preserving some temporal information. We also propose a statistical approach to learn and recognize individual gait properties from the limited training GEI templates.

In the next section, we introduce related work of human recognition by gait. The representation of GEI is introduced in Section 3. In Section 4, we propose two approaches for human recognition using GEI: direct GEI matching, and statistical GEI feature matching. In Section 5, we analyze the experimental results of the proposed human recognition approaches and compare them with the existing techniques. Section 6 concludes the paper.

2. Related Work

In recent years, various approaches have been proposed for human recognition by gait. These approaches can be divided into two categories: model-based approaches and model-free approaches.

2.1. Model-based Approaches

When people observe human walking patterns, they not only observe the global motion properties, but also interpret the structure of the human body and detect the motion patterns of local body parts. The structure of the human body is generally interpreted based on their prior knowledge. Model-based gait recognition approaches focus on recovering a structural model of human motion, and the gait patterns are then generated from the model parameters for recognition.

Niyogi and Adelson [14] make an initial attempt in a spatiotemporal (XYT) volume. They first find the bounding contours of the walker, and then fit a simplified stick model on them. A characteristic gait pattern in XYT is generated from the model parameters for recognition. Yoo et al. [19] estimate hip and knee angles from the body contour by linear regression analysis. Then trigonometric-polynomial interpolant functions are fitted to the angle sequences, and the parameters so-obtained are used for recognition. In Lee and Grimson's work [11], human silhouette is divided into local regions corresponding to different human body parts, and ellipses are fitted to each region to represent the human structure. Spatial and spectral features are extracted from these local regions for recognition and classification. Bhanu and Han [3] propose a kinematic-based approach to recognize individuals by gait. The 3D human walking parameters are estimated by performing a least squares fit of the 3D kinematic model to the 2D silhouette extracted from a monocular image sequence. Human gait signatures are generated by selecting features from the estimated parameters.

In these model-based approaches, the accuracy of human model reconstruction strongly depends on the quality of the extracted human silhouette. In the presence of noise, the estimated parameters may not be reliable. To obtain more reliable estimates, Tanawongsuwan and Bobick [17] reconstruct the human structure by tracking 3D sensors attached on fixed joint positions. However, their approach needs lots of human interaction which is not applicable in most surveillance applications.

2.2. Model-free Approaches

Model-free approaches make no attempt to recover a structural model of human motion. The features used for gait representation includes: moments of shape, height and stride/width, and other image/shape templates.

Moments of shape is one of the most commonly used gait features. Little and Boyd [12] describe the shape of human motion with a set of features derived from moments of a dense flow distribution. Shutler et al. [16] include velocity into the traditional moments to obtain the so-called velocity moments (VMs). A human motion image sequence can be represented as a single VM value with respect to a specific moment order instead of a sequence of traditional moment values for each frame. He and Debrunner's [7] approach detects a sequence of feature vectors based on Hu's moments of each motion segmented frame, and the individual is recognized from the feature vector sequence using hidden Markov models (HMMs).

BenAbdelkader et al. [2] use height, stride and cadence as features for human identification. Kale et al. [10] choose the width vector from the extracted silhouette as the representation of gait. Continuous HMMs are trained for each person and then used for gait recognition. In their later work [9], different gait features are further derived from the width vector and recognition is performed by a direct matching algorithm.

To avoid the feature extraction process which may reduce the reliability, Murase and Sakai [13] propose a template matching method to calculate the spatio-temporal correlation in a parametric eigenspace representation for gait recognition. Huang et al. [8] extend this approach by combining transformation based on canonical analysis, with eigenspace transformation for feature selection. BenAbdelkader et al. [1] compute the self-similarity plot by correlating each pair of aligned and scaled human silhouette in an image sequence. Normalized features are then generated from the similarity plots and used for gait recognition via eigenspace transformation.

As a direct template matching approach, Phillips et al. [15] measure the similarity between the gallery sequence and the probe sequence by computing the correlation of corresponding time-normalized frame pairs. Similarly, Collins et al. [5] first extract key frames from a sequence, and the similarity between two sequences is computed from normalized correlation. Tolliver and Collins [18] cluster human silhouettes/poses of each training sequence into k prototypical shapes. In the recognition procedure, the silhouettes in a testing sequence are also classified into k prototypical shapes which are compared to prototypical shapes of each training sequence for similarity measurement.

3. Gait Energy Image (GEI) Representation

In this paper, we only consider individual recognition by activity-specific human motion, i.e., regular human walking, which is used in most current approaches of individual recognition by gait.

3.1. Motivation

Regular human walking can be considered as cyclic motion where human motion repeats at a stable frequency. Some gait recognition approaches extract features from the correlation of all the frames in a walking sequence without considering their order [1, 8, 13]. Other approaches extract features from each frame and compose a feature sequence for the human walking sequence [2, 5, 7, 10, 9, 12, 15, 16, 18]. During the recognition procedure, they either match the extracted statistics from the feature sequence, or match the features between the corresponding pairs of frames in two sequences that are time-normalized with respect to their cycle lengths, respectively. The assumption here is that the order of poses in human walking cycles is the same, i.e., the limbs (arms and legs) move forward and backward in a similar way among normal people. The difference exists in the phase of poses in a walking cycle, the extend of limbs, and the shape of the torso, etc. As the order of poses in regular human walking is generally not considered in gait recognition approaches, it is possible to compose a spatio-temporal template in a single image instead of a ordered image sequences as usual.

3.2. Representation Construction

We use a silhouette extraction procedure and begin with the extracted binary silhouette sequences. The preprocess-

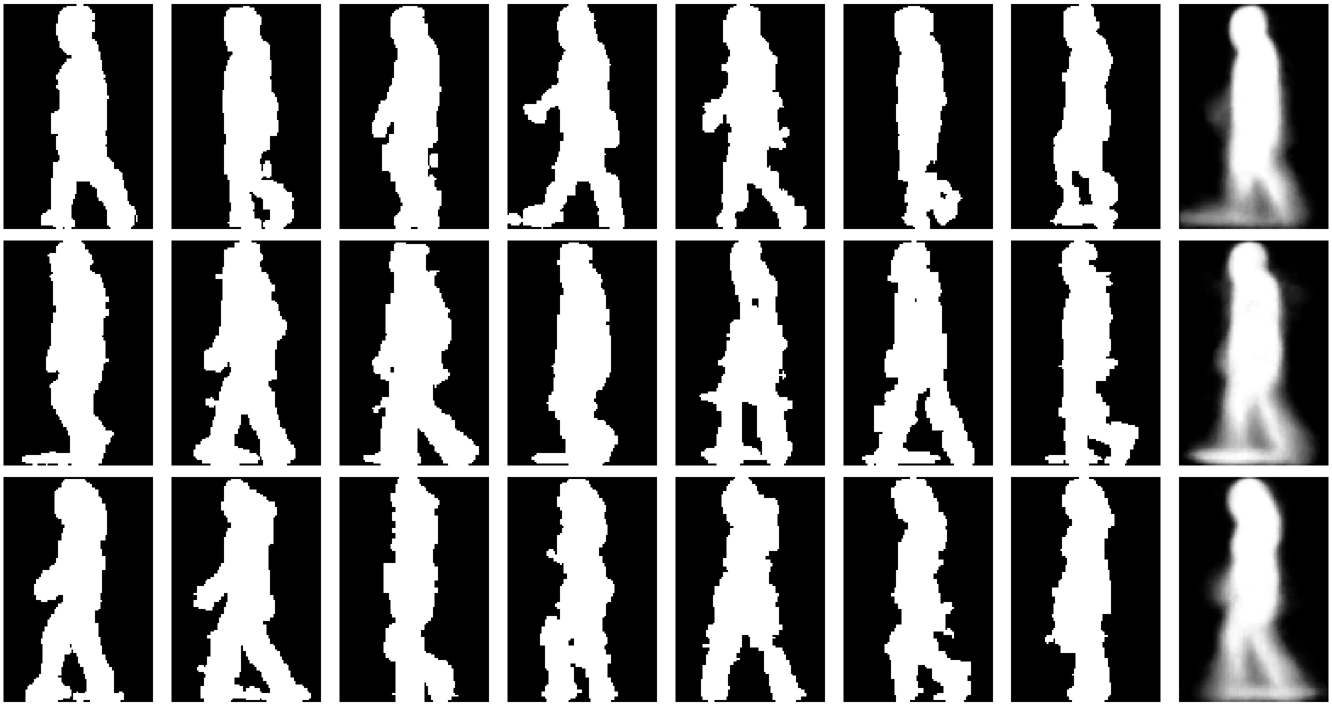


Figure 1. Examples of normalized and aligned silhouette frames in different human walking sequences. The rightmost image in each row is the average silhouette image over the whole sequence - Gait Energy Image (GEI).

ing procedure includes size normalization – fitting the silhouette height to the fixed image height, and sequential horizontal alignment – centering the upper half silhouette part with respect to the horizontal centroid. Figure 1 shows examples of preprocessed silhouette frames in different human walking sequences. The rightmost image in each row is the average silhouette image over the whole sequence. As expected, the average silhouette image reflects the major shapes of the human silhouettes and their changes over the sequence. A pixel with higher intensity value means that human body occurs more frequently at this position. Therefore, we refer to this average silhouette image as Gait Energy Image (GEI).

Given a size-normalized and horizontal-aligned human walking binary silhouette sequence $B(x, y, t)$, the grey-level GEI $G(x, y)$ is defined as follows

$$G(x,y) = \frac{1}{N}\sum_{t=1}^{N} B(x,y,t), \tag{1}$$

where N is the number of frames in complete cycles of the sequence, t is the frame number of the sequence, x and y are values in the 2D image coordinate.

3.3. Representation Justification

GEI has several advantages over the representation of binary silhouette sequence. As an average template, GEI is not sensitive to incidental silhouette errors in individual frames. The robustness could be further improved if we discard those pixels with the energy values lower than a threshold. Moreover, with such a 2D template, we do not need to divide the silhouette sequence into cycles and perform time normalization with respect to the cycle length. Therefore, the errors occuring in these procedures can be therefore avoid.

Compared with binary silhouette sequence, the information loss of GEI is obvious. For a specific pixel in GEI, we only know its intensity value, i.e., the frequency with which the human silhouette occurs at this position over the whole sequence. However, we might partly reconstruct the original silhouette sequence from the GEI according to the knowledge of regular human walking. For example, for a pixel near the outline of the leg area, it GEI value shows that silhouette occurs at this location in 20 frames out of 100 frames. Using the common sense, we know that 20 frames should be those frames where human stride instead of standing straight, if noise is not considered. Similarly, we can allocate the GEI values to most other leg/arm areas to corresponding frames in the silhouette sequence. In general, the energy changes in the torso and head area can be considered as noise. Although the knowledge is not enough to accurately allocate the GEI value of each pixel (i.e., the original silhouette sequence cannot be completely reconstructed), GEI still keeps the major shapes of human walking and reflects the major shape changes during walking. Actually, it is difficult to analyze how and in what degree the information loss affects the discriminating power of GEI as a template for individual recognition. We will evaluate this issue in the section of experimental results by comparing the recognition performance between GEI and binary silhouette sequence representations.

3.4. Relationship with MEI and MHI

Bobick and Davis [4] propose motion-energy image (MEI) and motion-history image (MHI) for human movement recognition. Both MEI and MHI are vector-image where the vector value at each pixel is a function of the motion properties at this location in an image sequence.

MEI is a binary image which represents where motion has occured in an image sequence:

$$E_\tau(x,y,t) = \cup_{i=0}^{\tau-1} D(x,y,t-i), \tag{2}$$

where $D(x,y,t)$ is a binary sequence indicating regions of motion, τ is the duration of time, t is the moment of time, x and y are values of 2D image coordinate. To represent a regular human walking sequence, if $D(x,y,t)$ is normalized and aligned as $B(x,y,t)$ in Equation (1), MEI $E_N(x,y,N)$ is the binary version of GEI $G(x,y)$.

MHI is a grey-level image which represents how motion in the image is moving:

$$H_\tau(x,y,t) = \begin{cases} \tau, & \text{if} \quad D(x,y,t)=1; \\ \max\{0, H_\tau(x,y,t-1)-1\}, & \text{otherwise.} \end{cases} \tag{3}$$

In general, both MEI and MHI are different motion representations compared to GEI. As regular human walking is a cyclic and highly self-occluded motion with a specific style, MEI and MHI are not suitable to represent regular human walking for individual recognition.

4. Human Recognition Using GEI Templates

Human walking sequences for training are limited in real surveillance applications. Because each sequence is represented as one GEI template, the training/gallery GEIs for each individual might limited to several or even one template(s). In this paper, we develop two approaches to recognize individuals from the limited templates.

4.1. Direct GEI Matching

One possible approach is recognizing individuals by measuring the similarity between the gallery (training) and probe (testing) templates. Given GEIs of two gait sequences, $G_g(x,y)$ and $G_p(x,y)$, their distance can be measured by calculating their normalized matching error:

$$D(G_g, G_p) = \frac{\sum_{x,y} |G_g(x,y) - G_p(x,y)|}{\sqrt{\sum_{x,y} G_g(x,y) \sum_{x,y} G_p(x,y)}}, \tag{4}$$

where $\sum_{x,y} |G_g(x,y) - G_p(x,y)|$ is the matching error between two GEIs, $\sum_{x,y} G_g(x,y)$ and $\sum_{x,y} G_p(x,y)$ are total energy in two GEIs, respectively.

This direct GEI matching approach is sensitive to distortion in silhouettes generated from image sequences that are recorded under different conditions. Recognition by learning may recover the inherent properties in training templates from an individual and therefore insensitive to such silhouette distortion. However, with one GEI template per individual, learning cannot be performed. Even with several templates per individual, if they are from similar conditions, the learned features may be overfit to the training templates.

4.2. Statistical GEI Feature Matching

In this section, we propose a statistical GEI feature matching approach for individual recognition from limited GEI templates. First, we generate new templates from the limited training templates according to a distortion analysis. Next, statistical features are learned from the expanded training templates by principal component analysis (PCA) to reduce the dimension of the template and multiple discriminant analysis (MDA) to achieve better class seperatability. The individual is recognized by the learned features. The system diagram of training and recognition procedure is shown in Figure 2.

4.2.1 Generating New Templates from Limited Training Templates

Various factors have effect on silhouettes extracted from the same person: shoe and clothing, walking surface, camera view, and shadow, etc. Shoe, surface and shadow affect the foot area of the silhouette. In addition, shoe and surface also change the human walking style. Clothing affects the shape of the silhouette. If the camera view changes slightly, there will be slight changes in silhouettes; if the camera view changes a lot, the extracted silhouettes may be totally different which may cause recognition to fail.

Among these factors, slight camera view changes may be neglected. The silhouette shape distortion incurred by the difference of clothing is irregular distortion, which occurs

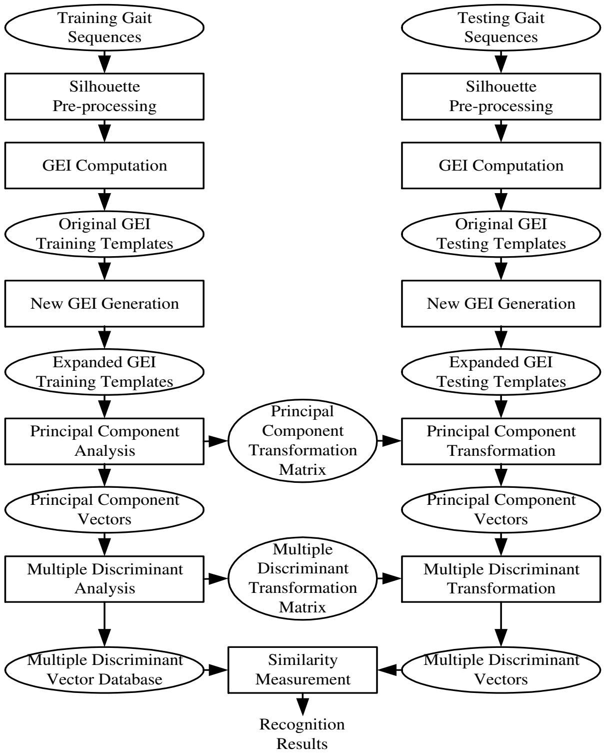

Training Gait Sequences
Silhouette Pre-processing
GEI Computation
Original GEI Training Templates
New GEI Generation
Expanded GEI Training Templates
Principal Component Analysis
Principal Component Vectors
Multiple Discriminant Analysis
Multiple Discriminant Vector Database
Testing Gait Sequences
Silhouette Pre-processing
GEI Computation
Original GEI Testing Templates
New GEI Generation
Expanded GEI Testing Templates
Principal Component Transformation
Principal Component Vectors
Multiple Discriminant Transformation
Multiple Discriminant Vectors
Principal Component Transformation Matrix
Multiple Discriminant Transformation Matrix
Similarity Measurement
Recognition Results

Figure 2. System diagram of individual recognition using the proposed statistical GEI feature matching approach.

in the upper body, lower body or both, and make body parts fatter or thinner. Thus, it is difficult to model this irregular distortion. Similarly, different shoes and walking surfaces incur global silhouette distortions which are also difficult to model. Now we consider the common distortion incurred by the difference of shoe, surface and shadow which generally occurs in the foot area of the silhouette. These distortions are local distortions which make the bottom part of the silhouette and GEI unreliable. If we generate new templates which are insensitive to the distortion in their bottom parts, the learned template properties will be insensitive to this kind of distortion.

The new GEI templates are generated as illustrated in Fig 3. First, we determine the range of the distortion area, e.g., n rows from the bottom row of the original GEI. Then, we cut a portion of the area from the bottom, and fit it to the original GEI size to obtain a new template. By repeating this step until reaching the upper row of the distortion area, we will obtain a series of new templates. The training templates expanded from the same original GEI have the same global shape properties but different bottom parts and different scales. Therefore, the learned features from the expanded training templates are insensitive to the common distortion by shadow, shoe and surface which occurs in the bottom part of GEI templates.

4.2.2 Learning Templates by Component Analysis and Discriminants

Once we obtain a series of training GEI templates for each individual, the problem of their excessive dimensionality occurs. To reduce their dimensionality, there are two classical approaches of finding effective linear transformations by combing features - Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA). As described in [6], PCA seeks a projection that best represents the data in a least square sense, while MDA seeks a projection that best separates the data in a least-square sense. Huang et al. [8] combine PCA and MDA to achieve the best data representation and the best class separability simultaneously. In this paper, the learning procedure follows this combination approach.

Given n d-dimensional training templates $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, PCA minimizes the criterion function

$$J_{d'} = \sum_{k=1}^{n} ||(\mathbf{m} + \sum_{i=1}^{d'} a_{ki}\mathbf{e}_i) - \mathbf{x}_k||^2, \tag{5}$$

where $d' < d$, $\mathbf{m} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$, and $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{d'}\}$ are a set of unit vectors. $J_{d'}$ is minimized when $\mathbf{e}_1$, $\mathbf{e}_2$, ..., and $\mathbf{e}_{d'}$ are the d' eigenvectors of the scatter matrix S having the largest eigenvalues, where

$$S = \sum_{k=1}^{n}(\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T. \tag{6}$$

The d'-dimensional principal component vector $\mathbf{y}_k$ is obtained from the d-dimensional GEI template $\mathbf{x}_k$ by multiplying the transformation matrix $[\mathbf{e}_1, ..., \mathbf{e}_{d'}]$:

$$\mathrm{y}_k = [a_1, ..., a_{d'}]^T = [\mathbf{e}_1, ..., \mathbf{e}_{d'}]^T \mathrm{x}_k, \quad k = 1, ..., n. \tag{7}$$

where n is the number of the expanded GEI templates from all people in the training dataset.

Although PCA finds components that are useful for representing data, there is no reason to assume that these components must be useful for discriminating between data in different classes because PCA does not consider the class label of training templates. Multiple discriminant analysis (MDA) seeks a projection that are efficient for discrimination. Suppose that the n d'-dimensional transformed training templates $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ belong to c classes. MDA seeks a transformation matrix W that in some sense maximizes the ratio of the between-class scatter S_B to the within-class scatter S_W:

$$J(W) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|W^T S_B W|}{|W^T S_W W|}. \tag{8}$$

The within-class scatter S_B is defined as

$$S_W = \sum_{i=1}^{c} S_i, \tag{9}$$

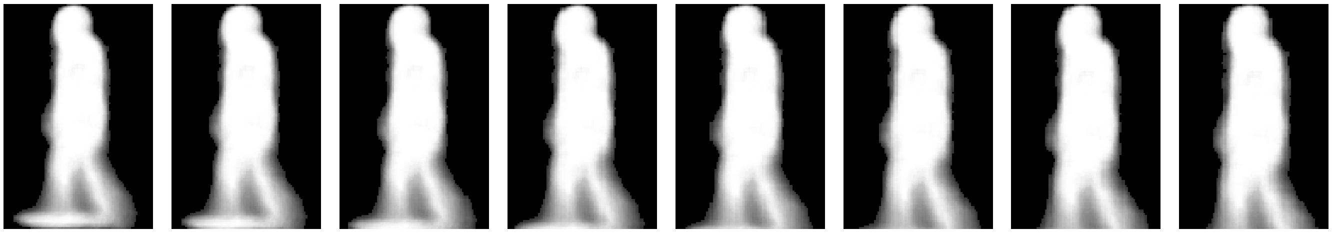


Figure 3. Examples of new GEI templates generated from the original template. The leftmost template is the original template, other templates are sequentially generated by cutting the bottom portion (2 rows in this example) of the previous template and fitting it to the original template size.

where

$$S_i = \sum_{\mathbf{y} \in \mathcal{D}_i} (\mathbf{y} - \mathbf{m}_i)(\mathbf{y} - \mathbf{m}_i)^T \tag{10}$$

and

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in \mathcal{D}_i} \mathbf{y}, \tag{11}$$

where $\mathcal{D}_i$ is the training template set that belongs to the ith class and n_i is the number of templates in $\mathcal{D}_i$. The within-class scatter S_B is defined as

$$S_B = \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \tag{12}$$

where

$$\mathbf{m} = \frac{1}{n} \sum_{\mathbf{y} \in \mathcal{D}} \mathbf{y}, \tag{13}$$

and $\mathcal{D}$ is the whole training template set. $J(W)$ is maximized when the columns of W are the generalized eigenvectors that correspond to the largest eigenvalues in

$$S_B \mathbf{w}_i = \lambda_i S_W \mathbf{w}_i. \tag{14}$$

There are no more than $c - 1$ nonzero eigenvalues, and the corresponding eigenvectors $\mathbf{v}_1$,...,$\mathbf{v}_{c-1}$ form transformation matrix. The $(c-1)$-dimensional multiple discriminant vector $\mathbf{z}_k$ is obtained from the d'-dimensional principal component vector $\mathbf{y}_k$ by multiplying the transformation matrix $[\mathbf{v}_1, ..., \mathbf{v}_{c-1}]$:

$$\mathrm{z}_k = [\mathbf{v}_1, ..., \mathbf{v}_{c-1}]^T \mathrm{y}_k, \quad k = 1, ..., n. \tag{15}$$

The obtained multiple discriminant vectors compose the feature database for individual recognition.

4.2.3 Individual Recognition

Given the GEI template $\tilde{\mathbf{x}}$ of a query gait sequence, a set of n_q templates $\{\tilde{\mathbf{x}}_1, ..., \tilde{\mathbf{x}}_{n_q}\}$ are generated according to the procedure described in Section 4.2.1. After the principal component transformation and multiple discriminant transformation, we obtain a set of feature vectors $\{\tilde{\mathbf{z}}_1, ..., \tilde{\mathbf{z}}_{n_q}\}$ for this test gait sequence. The feature distance between the query gait sequence and each class in the feature database can be given by the minimum distance between query and training feature vector pairs as follows

$$Distance_i = \min_{\mathbf{z} \in \mathcal{D}_i} \min_{j=1}^{n_q} \sum_{k=1}^{c-1} |\tilde{z}_{jk} - z_k|, \quad i = 1, ..., c. \tag{16}$$

After the distances for all classes are obtained, they are ranked in an ascending order where the class with the smallest distance is the best match of the query gait sequence.

5. Experimental Results

Our experiments are carried out on the USF HumanID May-2001 gait database. This database consists of 452 sequences from 74 persons walking in elliptical paths in front of the cameras. For each person, there are up to 5 covariates: viewpoints - Left/Right, shoe types - A/B, surface types - grass/concrete, carrying conditions - with/without a briefcase, and time and clothing. Seven experiments are designed for individual recognition as shown in Table 1. The gallery set contains 71 sequences. No sequence belongs to the same person in each individual data set.

Phillips et al. [15] proposed a baseline approach to extract human silhouettes and recognize individuals in this database. For comparison, they provide extracted silhouette data which can be found at the website http://marathon.csee.usf.edu/GaitBaseline/. Our experiments begin with these extracted binary silhouette data (parameterized version 1.7). The experimental results are shown in Table 2 and 3 as well as comparison with other approaches of individual recognition by gait. In these tables, rank1 means that only the first subject in the retrieval rank list is recognized as the same subject as the query subject, and rank5 means that the first five subjects are all recognized as the same subject as the query subject. The performance in these tables is the recognition rate under these two definitions.

Table 1. Seven experiments designed for individual recognition in USF HumanID database.

Experiment Label	Size of Probe Set	Difference between Gallery and Probe Sets
A	71	View
B	41	Shoe
C	41	View and Shoe
D	70	Surface
E	44	Surface and Shoe
F	70	Surface and View
G	44	Surface, Shoe and View

Table 2. Comparison of recognition performance of Rank 1 among different approaches on silhouette sequence version 1.7. (Legends: USF - direct frame shape matching [15]; DGEI - direct GEI matching, this paper; CMU - key frame shape matching [5]; SPS1/SPS2 - clustered frame shape matching with two criteria [18]; SGEI - statistical GEI feature matching, this paper.)

	USF	DGEI	CMU	SPS1	SPS2	SGEI
A	79%	99%	87%	82%	85%	90%
B	66%	83%	81%	66%	81%	90%
C	56%	73%	66%	54%	60%	73%
D	29%	18%	21%	20%	23%	41%
E	24%	14%	19%	18%	17%	40%
F	30%	11%	27%	21%	25%	27%
G	10%	10%	23%	21%	21%	38%

Table 3. Comparison of recognition performance of Rank 5 among different approaches on silhouette sequence version 1.7. (Same legend as in Table 2)

	USF	DGEI	CMU	SPS1	SPS2	SGEI
A	96%	100%	100%	98%	90%	99%
B	80%	93%	90%	90%	87%	93%
C	76%	93%	83%	81%	80%	93%
D	61%	55%	59%	46%	52%	68%
E	52%	52%	50%	43%	43%	69%
F	45%	47%	53%	46%	48%	58%
G	33%	52%	43%	43%	44%	60%

5.1. Recognition Results by Direct GEI Matching

To evaluate the effectiveness of GEI as a gait representation, we carry out experiments of individual recognition by direct matching between GEI templates according to the distance metric give by Equation (4). As we mentioned in Section 2.2, Phillips et al. [15] measure the similarity between the gallery sequence and the probe sequence by computing the correlation of corresponding time-normalized frame pairs. This approach can be viewed as a typical direct matching approach between regular gait silhouette sequences. We compare the recognition performance between their approach (USF) and our direct GEI matching approach (DGEI) as shown in Table 2 and 3.

The left part of Table 2 and 3 shows the recognition performance of USF and DGEI approaches. It is shown that our DEGI approach achieves much better results in experiments A-C. In these experiments, the difference between gallery and probe data exists in view, shoe or both, which incur little distortion in extracted silhouette. This means that GEI is less sensitive to this kind of distortion than regular gait silhouette sequence.

Although the rank1 performance of DGEI and USF are both not good in experiments D-G, our DEGI is worse than that of USF (See Table 2). The probe sets in experiments D-G have the common difference of surface with respect to the gallery set. As we discussed previously, the distortion incurred by surface difference is relatively high. For example, if the same person walks at different surface, the extracted silhouettes may have different shadows. In addition, the silhouette from a walking sequence on the grass surface may miss the bottom part of the feet because they could be covered by the grass. In this case, silhouette height normalization errors occur, and the silhouette so-obtained may have different scale with respect to the silhouette on other surfaces. It is shown that the GEI is sensitive to this kind of distortion with respect to the regular silhouette sequence. However, the rank5 performance of our DGEI is similar to that of USF in experiments D-G (See Table 3). This shows that GEI is competitive with regular silhouette sequence because the rank1 results are not reliable and more ranked subjects should be considered in these experiments. Another reason of the rank1 worse performance of DGEI (See Table 3) is that silhouettes of version 1.7 are not well-aligned.

5.2. Recognition Results by Statistical GEI Feature Matching

Table 2 and 3 show that our individual recognition approach by statistical GEI feature matching (SGEI) achieves better recognition results than DGEI in the experiments with large silhouette distortion, i.e., D-G. In other experiments with small silhouette distortion, the performance of SGEI is better than that of DGEI in experiments B and C, but slightly worse in experiments A. Thus SGEI slightly sacrifices the performance in experiments with small silhouette distortion while improving the performance in experiments with large silhouette distortion with respect to DGEI.

We also compare the performance of SGEI with other approaches published in [15, 5, 18] in Table 2 and 3. It is shown that SGEI achieves better or equivalent recognition performance than other approaches in all experiments.

6. Conclusions

In this paper, a new spatio-temporal gait representation, called Gait Energy Image (GEI), is proposed for individual recognition by gait. Unlike other gait representation which considers gait as a sequence of templates (poses), GEI represents a human motion sequence in a single image while preserving temporal information. To overcome the limitation of training templates, we generate a series of new GEI templates by analyzing the human silhouette distortion under various conditions. Principal component analysis and multiple discriminant analysis are used for learning features from the expanded GEI training templates. Recognition is then carried out based on the learned features. Experimental results show that (a) GEI is an effective and efficient gait representation which is insensitive to incidental silhouette errors in individual frames, and (b) the proposed recognition approach achieves highly competitive performance with respect to the published gait recognition approaches.

Acknowledgment

This work was supported in part by grants F49620-02-1-0315 and DAAD19-01-0357; the contents and information do not necessarily reflect the position or policy of U.S. Government. The authors would like to thank Dr. Sudeep Sarkar (University of Southern Florida) and Dr. Patrick Grother (National Institute of Standards and Technology) for their help in providing the HumanID dataset and silhouette data.

References

[1] C. BenAbdelkader, R. Cutler, and L. Davis. Motion-based recognition of people in eigengait space. in *Proc. International Conference on Automatic Face and Gesture Recognition*, pages 254–259, May 2002.

[2] C. BenAbdelkader, R. Cutler, and L. Davis. Person identification using automatic height and stride estimation. in *Proc. International Conference on Pattern Recognition*, 4:377–380, 2002.

[3] B. Bhanu and J. Han. Individual recognition by kinematic-based gait analysis. in *Proc. International Conference on Pattern Recognition*, 3:343–346, 2002.

[4] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, March 2001.

[5] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 351–356, May 2002.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Willy & Sons, 2000.

[7] Q. He and C. Debrunner. Individual recognition from periodic activity using hidden markov models. in *Proc. IEEE Workshop on Human Motion*, pages 47–52, 2000.

[8] P. Huang, C. Harris, and M. Nixon. Recognizing humans by gait via parameteric canonical space. *Artificial Intelligence in Engineering*, 13:359–366, 1999.

[9] A. Kale, N. Cuntoor, B. Yegnanarayana, A. Rajagopalan, and R. Chellappa. Gait analysis for human identification. in *Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 2688*, pages 706–714, 2003.

[10] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger. Gait-based recognition of humans using continuous hmms. in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 321–326, May 2002.

[11] L. Lee and W. Grimson. Gait analysis for recognition and classification. in *Proc. 5th International Conference on Automatic Face and Gesture Recognition*, pages 148–155, May 2002.

[12] J. Little and J. Boyd. Recognizing people by their gait: the shape of motion. *Videre: Journal of Computer Vision Research*, 1(2):1–32, 1998.

[13] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognition Letters*, 17(2):155–62, 1996.

[14] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. in *Proc. IEEE Conference on CVPR*, pages 469–474, 1994.

[15] P. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer. The gait identification challenge problem: data sets and baseline algorithm. in *Proc. IEEE International Conference on Pattern Recognition*, 1:385–388, 2002.

[16] J. Shutler, M. Nixon, and C. Harris. Statistical gait recognition via velocity moments. in *Proc. IEE Colloquium on Visual Biometrics*, pages 10/1–10/5, March 2000.

[17] R. Tanawongsuwan and A. Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2:726–731, 2001.

[18] D. Tolliver and R. Collins. Gait shape estimation for identification. in *Proc. 4th International Conference on Audio- and Video-Based Biometric Person Authentication, LNCS 2688*, pages 734–742, 2003.

[19] J.-H. Yoo, M. Nixon, and C. Harris. Model-driven statistical analysis of human gait motion. in *Proc. IEEE International Conference on Image Processing*, 1:285–288, Sept. 2002.

Model-Based Gait Enrolment in Real-World Imagery

David K Wagg and Mark S Nixon

Department of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK
dkw02r | msn@ecs.soton.ac.uk

Abstract

We present a model-based approach to gait extraction that is capable of reliable operation on real-world imagery. Hierarchies of shape and motion are employed to yield relatively modest computational demands, avoiding the high-dimensional search spaces associated with complex models. Anatomical data is used to generate shape models consistent with normal human body proportions. Mean gait data is used to create prototype gait motion models, which are adapted to fit individual subjects.

Accuracy is evaluated on subjects filmed from a fronto-parallel view in controlled laboratory conditions, for which some gait parameters are known. We further show that comparable performance is attained in outdoor conditions. As such, we describe a new approach to enrolment for gait recognition technologies, allowing reliable subject gait extraction in real-world imagery.

1. Introduction

Gait may be defined as the individual pattern of movement produced as a person walks. This pattern is sufficiently unique for each individual to be employed as a biometric [Winter91, Nixon99]. Gait analysis is usable from a distance and does not require the subject to be aware of or cooperate with its use, making it particularly valuable in surveillance, or other applications where non-contact operation is required.

This field is currently dominated by face recognition, supported by the role of facial features in the human recognition process. However, gait is more difficult to obscure or disguise, and can be measured from a much wider range of viewpoints. Gait is also more robust with respect to occlusion and variations in illumination, as a gait signature is spatio-temporal rather than a purely spatial measure.

Gait may be best employed in combination with other biometrics, with facial features being an obvious choice. Most approaches to face recognition require a relatively constrained frontal viewpoint, and gait could be employed as a back-up strategy when the subject's face is not visible. Alternatively, multiple cameras could be employed to combine face and gait features, improving overall recognition performance [Shakhnarovich01].

However, enrolment is a more difficult problem for gait, particularly when enrolment conditions cannot be controlled (for example, when enrolling a subject from CCTV footage). Gait enrolment requires the extraction of limb dynamics over a period of time, ideally capturing at least one full gait cycle. In uncontrolled capture conditions, it is likely that other objects will interfere with and occlude the subject; in addition gait is partially self-occluding, as one leg passes in front of the other. To successfully resolve this problem, extraction methodologies must be highly robust to noise and occlusions.

Many existing approaches to gait enrolment are data-driven, typically using the person's silhouette or features derived features from it as a basis for recognition [BenAbdelkader02, Collins02, Huang99, Johnson01, Kale03, Lee02, Phillips02]. This methodology has many advantages, chiefly of speed and simplicity, but has the disadvantage that silhouette dynamics are only indirectly linked to gait dynamics. Noise, occlusions and variations in clothing will all affect silhouette dynamics; it is unclear how a silhouette-based feature set could be normalised for these factors.

Model-based approaches overcome these weaknesses by incorporating knowledge of the shape and dynamics of human gait into the extraction process [Cunado03, Meyer98, Yam02]. The use of a model ensures that only image data corresponding to allowable human shape and motion is extracted, reducing the impact of noise. It also means that gait dynamics are extracted directly by determining joint positions, rather than inferring dynamics from other measures. A model-based approach also has the potential for more general applications, such as animation, user interfaces or model-based coding [Gavrila99].

However, the use of a parametric model introduces its own problems. Success in recognition is dependent on the gait signature being sufficiently complex to incorporate individual variation across the subject population, so that a given subject can be distinguished from all the other subjects under test. As gait is dependent on a large number of parameters (such as joint angles and body segment sizes), this requirement leads to complex models with many free parameters. Finding the best fitting model for the subject thereby necessitates searching a high-dimensional parameter space, with correspondingly high computational requirements.

Most early approaches dealt with this problem by severely limiting model complexity; later solutions have improved on this situation somewhat. [Nash 98] employs a genetic algorithm to cope with the high computational demands, but due to its reliance on stochastic processes this strategy cannot guarantee an optimal model fit. [Lappas02] introduces the dynamic velocity Hough transform, which applies dynamic programming to find an optimal object trajectory using structural evidence and smoothness of motion constraints. However, under this formulation it is difficult to apply parametric motion constraints (such as pendular limb motion).

To reduce the computational requirements of a model-based approach, we employ a model hierarchy composed of shape and motion components.

A velocity filtering algorithm is employed to determine the bulk motion of the person independently of shape parameters. Using this motion information we form a global temporal accumulation describing the person's average shape over the gait sequence. This accumulation is used to robustly estimate the size and shape of the person's body segments, using ellipses for the head and torso and two pairs of lines for each leg, applying anatomical constraints to reduce matching errors. Using this initialisation we can estimate the dominant gait frequency via a measurement of edge strength about the lower leg region over time. Leg motion is estimated by fitting prototype gait curves collected from a clinical gait study, stretched or compressed to fit the subject's gait frequency and hip rotational amplitude.

Our approach currently assumes a single subject moving at a constant speed, fronto-parallel against a cluttered background. However, this approach could be generalised to an arbitrary viewpoint.

We show that this methodology provides a good initial model fit suitable for further adaptation, and is capable of performance in noisy real-world conditions similar to that in controlled laboratory conditions.

2. Gait Signature Extraction

2.1. Bulk Motion Estimation

We may consider the motion of a person in normal gait to be composed of many separate motion components, forming a hierarchy according to the total pixel displacement they are responsible for. At the top of this hierarchy is the person's velocity in the horizontal plane, as a person will move with approximately constant velocity during normal gait (changes in velocity may also distort their gait signature, further justifying this assumption). The second level of the hierarchy is articulated motion; we may consider a third level to be object deformations (for example due to clothing or camera distortion), but this level of detail is considered unnecessary for our current purposes.

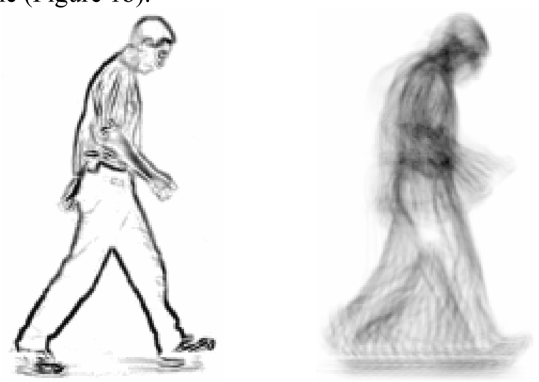
Image data is pre-processed (Figure 1a) using a Gaussian averaging filter for noise suppression, followed by Sobel edge detection and background subtraction (the background is computed by a temporal median of neighbouring frames). This removes all static objects, leaving only edges belonging to moving objects. The extraction process does not require binary edge data, which means that error-prone thresholding can be avoided.

Using a velocity filtering algorithm it is possible to determine object motion independently of shape. This algorithm effectively performs the same global temporal accumulation as the velocity Hough transform [Nash97], but without shape specificity:

$$A_v(i,j) = \sum_{n=0}^{N} I_n\left(i + v\left(\frac{N}{2} - n\right), j\right) \qquad (1)$$

where A_v is the accumulation for velocity v (in pixels per frame), I_n is the image intensity function at frame n, i and j are coordinate indices and N is the number of frames in the gait sequence.

This algorithm sorts objects in the scene according to their velocity and starting position, producing an accumulation for each possible object velocity. Each object's contribution to an accumulation is dependent on its edge strength, the number of frames it is in view of the camera and how close its velocity is to the accumulation velocity. This global averaging process means that objects in each accumulation are relatively unaffected by other objects, greatly reducing the problems associated with objects merging and splitting. At the correct accumulation velocity for an object, edges from each frame will accumulate to a single area, producing an average shape outline (Figure 1b).



(a) Section of pre-processed image data (b) Global temporal accumulation

Figure 1: Motion estimation by temporal accumulation

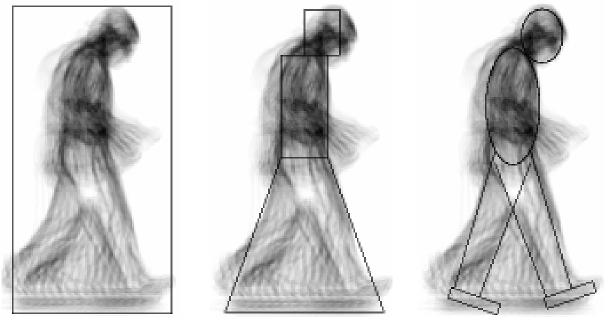
Each moving object in the scene appears as a peak in a plot of maximal accumulation intensity against velocity. Assuming that the person is the most significant moving object in the scene, their velocity can be inferred by selecting the highest peak in this plot (this assumption holds true for most current gait databases). If there are other more significant objects within the scene moving at a similar velocity, we must apply some knowledge of the person's shape to distinguish them from the other objects.

Noting that Equation 1 simply shifts and accumulates each frame, we can improve computational efficiency by first run-length encoding the input data. This representation is shift-invariant, and as runs of zero magnitude edge strength can simply be discarded, this reduces the order of the algorithm to O($V \cdot E \cdot N$), where V is the number of possible velocities, E is the mean number of edge points in each frame and N is the number of frames in the gait sequence. Further performance improvement can be accrued by downsampling input frames and applying a coarse-to-fine velocity search strategy.

2.2. Shape Estimation

The temporal accumulation computed during the bulk motion estimation stage forms an average global view of the person's shape. Parameters that do not change over the course of the gait cycle can therefore be determined from the temporal accumulation; as it is robust with respect to noise and occlusion, static parameters can be estimated with confidence.

The size and proportions of the person are estimated in a hierarchical fashion using anatomical constraints, derived from data published in [Winter90]. A region-growing algorithm is first applied to find all edges belonging to the person. This algorithm is initialised at the peak point in the accumulation, and an aspect-ratio constrained rectangular region is expanded about the point until all significant edges have been encompassed (Figure 2a).



(a) Region expansion

(b) Coarse segmentation

(c) Final shape estimate

Figure 2: Shape extraction hierarchy

Using this initialisation the approximate height of the person is estimated, using a fixed body segmentation based on mean anatomical proportions (Figure 2b). The final shape model (Figure 2c) consists of two ellipses for the head and torso, two rectangles for the feet and two pairs of lines for each leg. The parameters describing the head and torso are determined by template matching within the locality of the initial segmentation, constrained by mean anatomical proportions. The leg and foot shape parameters are computed as a fixed proportion of the subject's height and torso width, again based on mean anatomical data.

Note that although all shape dynamics are lost in the accumulation process, it is still possible to estimate the amplitude of hip rotation, which may be used to aid articulated motion estimation.

2.3. Articulated Motion Estimation

The motion of the leg during normal gait is periodic, and may be approximately modelled by a single sinusoid [Cunado03]. Applying this assumption, we can estimate a person's gait cycle frequency by measuring edge strength within the outer region of their legs, throughout the gait sequence (Figure 3).

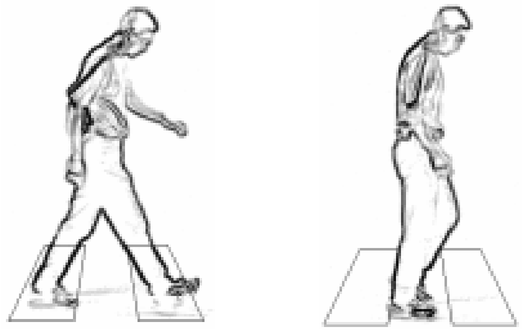
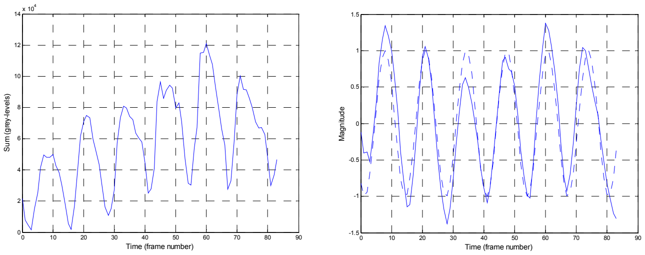


Figure 3: Gait cycle frequency estimation using within-region edge strength measurements

These measurements form a signal with approximately sinusoidal shape, distorted and contaminated by noise due to varying illumination, occlusion and motion estimation errors. Figure 4a depicts this signal for an example outdoor sequence:



(a) Measured signal

(b) Normalised data and sinusoidal fit

Figure 4: Gait cycle frequency estimation

Signal distortions are corrected by using low-order polynomials to model variation in the mean level of the sinusoid (numerator of Equation 2), and local variations in sinusoid magnitude (denominator of Equation 2):

$$S_n = \frac{S - p(S)}{p(\,|\,S - p(S)\,|\,)} \qquad (2)$$

where S_n is the normalised signal, S is the original signal and $p(x)$ denotes the best 2nd-order polynomial fit to signal x, computed by least-squares regression.

Frequency estimation is performed by fitting a fixed-amplitude sinusoid to the data, selecting the frequency and phase that minimises squared error (Figure 4b).

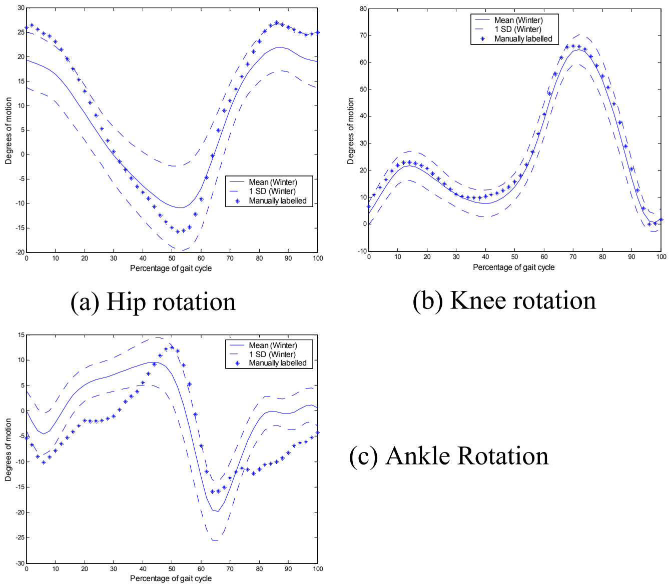
This frequency information can be applied directly using sinusoidal joint rotation models [Cunado03, Yam02]. A single sinusoid is adequate to approximately model the rotation of the hip and knee joints:

$$\theta_h(t) = A_h \sin\left(wt + \varphi_h\right) + \psi_h \qquad (3)$$

$$\theta_k(t) = A_k \sin\left(wt + \varphi_h + \varphi_k\right) + \psi_k \qquad (4)$$

where $\theta_h(t)$ and $\theta_k(t)$ are the respective hip and knee joint rotations (measured relative to the vertical axis) at time t, A_h and A_k are the joint rotational amplitudes, w is the gait cycle frequency (in radians per frame), ϕ_h is the starting hip joint phase, ϕ_k is a constant phase offset, ψ_h and ψ_k are constant amplitude offsets.

However, accuracy can be improved by more closely modelling human gait. Clinical gait studies have quantitatively measured the pattern of movement produced as people walk, by attaching markers to each joint. Mean gait patterns from [Winter91] were used to produce prototypical rotation models for the hip, knee and ankle joints. Figure 5 shows these models, together with joint angles manually extracted from a sequence in the Southampton HiD database [Shutler02]. Note that by clinical convention rotations are measured in degrees of motion, as opposed to rotation relative to the vertical axis.



(a) Hip rotation

(b) Knee rotation

(c) Ankle Rotation

Figure 5: Mean joint rotation patterns

This comparison suggests that the mean rotation models for the hip and knee match well to a typical subject. Ankle rotation is not such a good match, as the subjects in the clinical study were barefoot, as opposed to a typical subject who will be wearing shoes. However, the mean ankle rotation model still provides a better basis than a simple sinusoidal model would. The motion of the pelvis is not modelled at this point; the positions of the hip joints are assumed to coincide, remaining at the same level throughout the gait sequence.

The discrete Fourier transform (DFT) of each model is computed, creating continuous representations of the shape of the models. To match the subject's gait, the DFT models are scaled to match the subject's estimated gait cycle frequency and hip amplitude. Cycle phase is estimated by temporally matching leg templates to edge strength over the whole sequence, selecting the phase that maximises template correlation. Matching globally in this fashion increases resistance to noise, and can be performed quickly when only one search parameter is required.

Finally, the vertical oscillation of the subject's upper body is modelled by a single sinusoid with parameters proportional to the subject's height and gait motion:

$$Y(t) = A_y \sin 2\left(wt + \varphi_h + {}^{\pi}\!/_{8}\right) + \psi_y \quad (5)$$

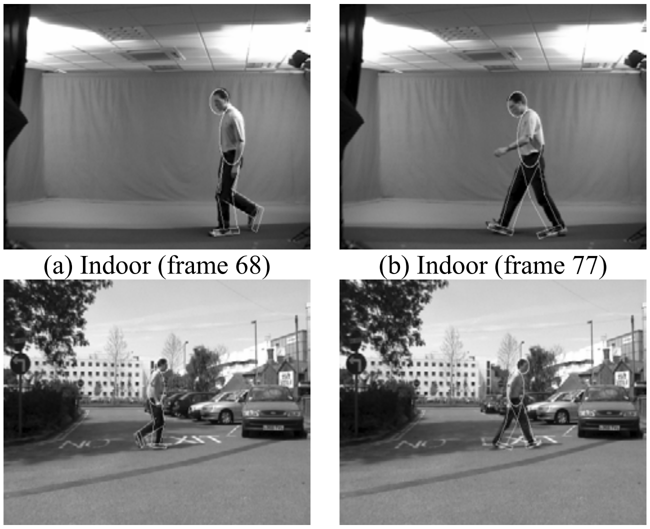
where $Y(t)$ is the y-coordinate of the torso at time t, A_y is the amplitude of oscillation, w is the gait cycle frequency, ϕ_h is the starting hip joint phase and ψ_y is the centre of oscillation.

The joint positions extracted by this process only approximate the true joint positions (the estimation process effectively assumes average gait motion, or no individuality). However, these positions form a strong basis for further model adaptation, which would make recognition possible.

3. Results

The performance of the gait extraction process was evaluated on sequences of two subjects from the Southampton HiD database [Shutler02]. Each subject was filmed from a fronto-parallel viewpoint, in controlled laboratory conditions and in noisy outdoor conditions, allowing the noise-resistance to be tested in isolation from other variables. The database is encoded in Digital Video (DV) format at a resolution of 720x576 pixels, recorded at a rate of 25 frames per second. Each sequence typically consists of 80-100 frames, or around 3 full gait cycles.

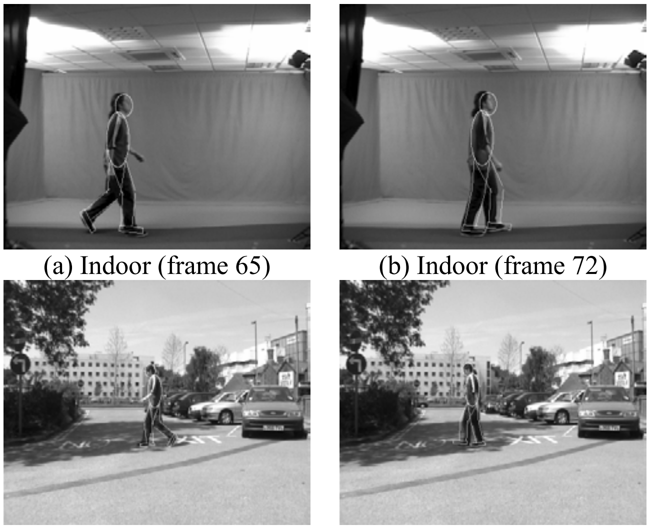
The extraction process is fast, with approximately 75% of the total processing time taken up by pre-processing. A 2.4GHz Pentium 4-based PC was used for all testing, requiring approximately 30 seconds processing time for each sequence. Figures 6 and 7 give some examples of the extraction process, showing good overall performance, especially on the outdoor data. Note that there is some error evident in shape estimation, and also some error caused by the assumption that the left and right hip joints coincide.



(a) Indoor (frame 68) (b) Indoor (frame 77)

(c) Outdoor (frame 39) (d) Outdoor (frame 47)

Figure 6: Sample extraction results for subject 013



(a) Indoor (frame 65) (b) Indoor (frame 72)

(c) Outdoor (frame 49) (d) Outdoor (frame 56)

Figure 7: Sample extraction results for subject 014

The set-up of the indoor data allows an approximation to ground truth to be made by chroma-key extraction of the subject's silhouette [Shutler02]. From this silhouette data the frame numbers at which the subject's heel strikes the floor are recorded, so that a comparison can be made with the automatically extracted result. The heel-strike frames were estimated from the automatic extraction by finding the knee rotation minima over the sequence.

Although this does not yield an exact measure of the extraction performance, this evaluation can be performed automatically on a large number of sequences. Table 1 shows the results of this evaluation for 56 indoor test sequences split equally over four subjects:

Table 1: Extraction performance under controlled conditions – RMS error in predicted heel-strike frames

Subject	Mean	Standard Deviation
013 (M)	0.933	0.236
014 (M)	0.954	0.458
033 (F)	0.741	0.209
037 (F)	0.979	0.363

The mean error in estimating the point of heel-strikes is around ±1 frame for both subjects, comparable to typical human labelling error. This is an encouraging result, demonstrating that we can successfully track the motion of the subject's legs in relatively clean indoor conditions. To demonstrate robustness, the extraction process was repeated on outdoor data, totalling 64 sequences of the same four subjects. As no ground truth data is available for the outdoor dataset, extraction performance is estimated by comparing the gait cycle period extracted from the outdoor data to that of the indoor data (Figures 8 and 9):

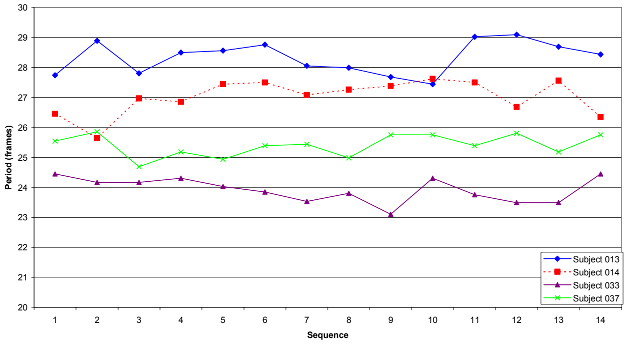


Figure 8: Period extraction for indoor data

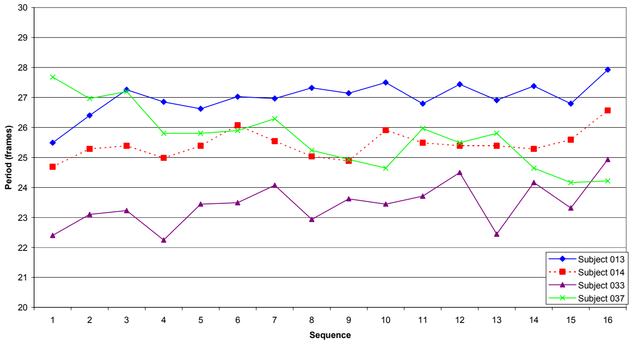


Figure 9: Period extraction for outdoor data

The extracted period is generally consistent between different gait sequences for each subject. Some of the subjects exhibit a reduced gait period on the outdoor dataset, indicating increased cadence. This may be due to the walking surface, or possibly because the subjects do not have a limited walking track in the outdoor dataset. However, even with only one gait parameter most of the subjects can be distinguished from one another.

For a more detailed view of performance, one indoor and one outdoor sequence was manually labelled for each test subject. The positions of the hip, knee and ankle joints were recorded, for comparison against the automatically extracted joint positions. The error is measured by a Euclidean distance metric, normalised to a percentage of the height of the subject. This error is given for a mean gait cycle, averaged over the sequence.

Figure 10 shows the errors measured at each joint position for subject 013 from the Southampton HiD database. Note that some error is expected of the human labelling, estimated at around 1% of subject height (the height of a subject is typically around 300 pixels on the indoor data or 200 pixels on the outdoor data).

This comparison shows that the additional increase in error when moving from controlled laboratory conditions to outdoor conditions is relatively small. It also shows that the additional complexity imposed by the use of mean gait rotation models is justified, resulting in a significant reduction in error over the sinusoidal models (Equations 3 and 4). The motion produced by these models is noticeably more natural in appearance to the human observer, suggesting that further improvement in performance is possible.

4. Conclusions

We have presented a new model-based gait enrolment technique to allow the use of gait analysis on real-world imagery. A model hierarchy of shape and motion keep the computational requirements of this approach to a minimum, while retaining the well-known robustness of a model-based approach.

Anatomical data and mean gait data is applied to produce shape and motion models adhering to known human proportions and gait dynamics, minimising the modelling error in this approach.

We have shown that we can reliably locate joint positions for the purposes of gait analysis in real-world imagery, with only a small loss in accuracy compared to controlled laboratory conditions. Future work will extend this approach by adapting the mean gait models to match each individual, so that recognition may be performed on the gait parameterisation thus obtained.

Acknowledgements
We gratefully acknowledge partial support by the European Research Office of the US Army under Contract No. N68171-01-C-9002.

References

[BenAbdelkader02] C BenAbdelkader, R Cutler and L Davis. "Stride and Cadence as a Biometric in Automatic Person Identification and Verification." *Proc. FGR*, pp. 372-377, 2002.

[Collins02] R T Collins, R Gross and J Shi. "Silhouette-based Human Identification from Body Shape and Gait." *Proc. FGR*, pp. 351-356, 2002.

[Cunado03] D Cunado, M S Nixon and J N Carter. "Automatic Extraction and Description of Human Gait Models for Recognition Purposes." *CVIU*, **90** (1), pp. 1-41, 2003.

[Gavrila99] D M Gavrila. "The Visual Analysis of Human Movement: A Survey." *CVIU*, **73** (1), pp. 82-98, 1999.

[Huang99] P. S. Huang, C. J. Harris and M. S. Nixon. "Recognizing Humans by Gait via Parametric Canonical Space." *Artificial Intelligence in Engineering*, **13** (4), pp. 359-366, 1999.

[Johnson01] A Y Johnson and A F Bobick. "A Multi-View Method for Gait Recognition Using Static Body Parameters." *Proc. AVBPA*, pp. 301-311, 2001.

[Kale03] A Kale, N Cuntoor, B Yegnanarayana, A N Rajagopolan and R Chellappa. "Gait Analysis for Human Identification." *Proc. AVBPA*, 2003.

[Lee02] l Lee and W E L Grimson. "Gait Analysis for Recognition and Classification." *Proc. FGR*, pp. 155-162, 2002.

[Meyer98] D Meyer, J Posl and H Niemann. "Gait Classification with HMMs for Trajectories of Body Parts Extracted by Mixture Densities." *Proc. BMVC*, pp. 459-468, 1998.

[Nash97] J M Nash, J N Carter and M S Nixon. "Dynamic Feature Extraction via the Velocity Hough Transform." *Pattern Recognition Letters*, **18**, pp. 1035–1047, 1997.

[Nash98] J M Nash, J N Carter and M S Nixon. "Extraction of Moving Articulated-Objects by Evidence Gathering." *Proc. BMVC*, pp. 609-618, 1998.

[Phillips02] P J Phillips, S Sarkar, I Robledo, P Grother and K Bowyer. "The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm." *Proc. FGR*, pp. 137-142, 2002.

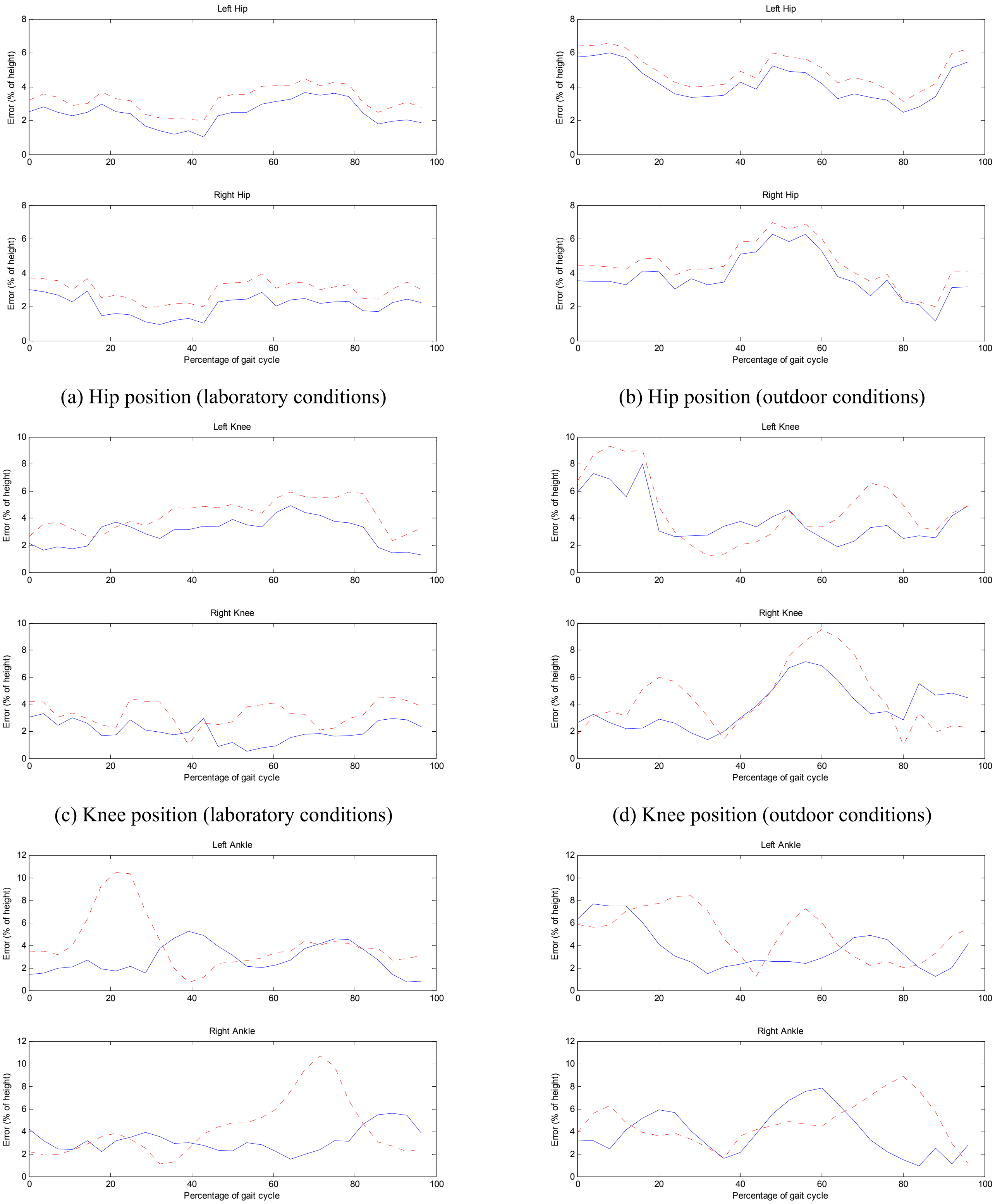
[Shakhnarovich01] G Shakhnarovich, L Lee and T Darrell. "Integrated Face and Gait Recognition from Multiple Views". *Proc. CVPR*, pp. 439-446, 2001.

[Shutler02] J D Shutler, M G Grant, M S Nixon and J N Carter. "On a Large Sequence-based Human Gait Database." *Proc. RASC*, pp. 66-71, 2002.

[Winter90] D A Winter. "Biomechanics and Motor Control of Human Movement (2nd Edition)." *John Wiley and Sons*, 1990.

[Winter91] D A Winter. "The Biomechanics and Motor Control of Human Gait: Normal, Elderly and Pathological." *University of Waterloo press, Ontario*. 1991.

[Yam02] C Yam, M S Nixon and J N Carter. "On the Relationship of Human Walking and Running: Automatic Person Identification by Gait." *Proc. ICPR*, pp. 287-290, 2002.



(a) Hip position (laboratory conditions)

(b) Hip position (outdoor conditions)

(c) Knee position (laboratory conditions)

(d) Knee position (outdoor conditions)

(e) Ankle position (laboratory conditions)

(f) Ankle position (outdoor conditions)

Figure 10: Error in automatically extracted joint positions from manually labelled positions (subject 013).
Solid line – mean gait models Dotted line – sinusoid model

Registration of Elastic Deformations of Fingerprint Images with Automatic Finding of Correspondences

Oleg Ushmaev, Sergey Novikov
BioLink Technologies International, Inc.
http://www.BioLinkUSA.com
E-mail: Oushmaev@BioLinkUSA.com, Snovikov@BioLinkUSA.com

Abstract

We elaborate the application of elastic deformation theory to fingerprint recognition. We propose the analytic model of elastic fingerprint deformations and its application to real fingerprint images. Also we carry out the statistical analysis of deformations of fingerprint images appearing in real applications.

1. Introduction

At the moment increasing power of computers facilitates the replacement of laborious manual fingerprint classification and matching methods by automatic fingerprint identification systems (AFIS) and automatic fingerprint authentication systems (AFAS).

AFIS [1],[2],[3] are most widely used (mainly for criminal search and related tasks) and usually have a fingerprint as an input data and the output is the list of identities of persons that could have the fingerprint given and a score for each identity indicating the similarity between two fingerprints. Such systems compare an input image with multiple of records in database.

AFAS [4], also referred as verification systems, are used in biometrics (detection of human identity by biological features) for access control and other civil applications. The input data in such systems are an identity and a fingerprint image, the output is an answer of Yes or No indicating whether the input image belongs to the person whose identity is provided.

In these applications there are four possible outcomes:

(1) an authorized person is accepted,
(2) an authorized person is rejected,
(3) an unauthorized person is accepted,
(4) an unauthorized person is rejected.

The rates of cases 2 and 4, which are called False Rejection Rate (FRR or FNMR, what means false non-match rate) and False Acceptance Rate (FAR or FMR, false match rate), are standardized metrics of identification accuracy of biometric systems [5]. The theoretical limits of FAR for different biometrics could be found in [6],[7].

Recently there appeared a scope of problems concerning the submission of a certain ID document (passport, driver license etc.) to one and only one particular person, thus a number of so called “civil ID” systems were created [8],[9],[10]. Usually such systems are required to have very little FAR.

There are a number of factors sufficiently raising the level of FAR. They can be divided into two major parts: poor quality of fingerprint images and human factor (improper applications). Quality of fingerprints can be sufficiently improved by different kind of filtering procedures. Improper applications often lead to full lost of information or to appearance of different distortions and deformations that have nothing in common with random noises and cannot be filtered (example of moderately deformed fingerprint images is presented in figures 1 and 2). In spite of existence of developed theory of elastic deformations, it is rarely applied to the real-time systems due to computational complexity.

There are different approaches to registration of elastic deformations. The way suggested by A.M. Bazen and S.H. Gerez [11] is based on the thin-plate spline (TPS) models, firstly applied to biological objects by F.L. Bookstein [12]. This method requires determining correspondent points in two compared images (matching point) and it suffers from the lack of precision in case of few matching points. Modifications of TPS (approximate thin-plate splines and radial based function splines) were introduced by M. Fornefett, K. Rohr and H. Stiehl [13],[14]. They consider deformations of biological tissues. But this way also requires many matching points (more then 100) what is virtually impossible in fingerprint applications, because number of minutiae in fingerprint image rarely exceeds

50. This fact makes TPS and its variants hardly applicable to fingerprint deformations registration.

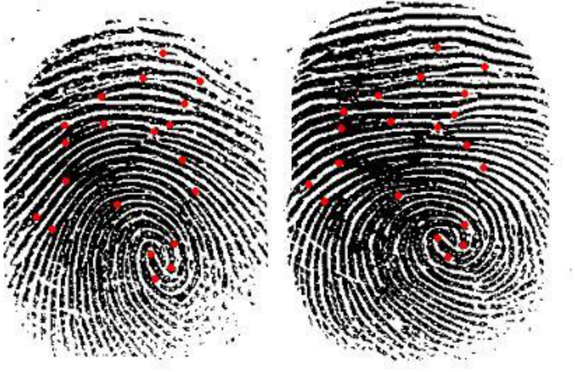


Figure 1. Pair of moderately deformed images

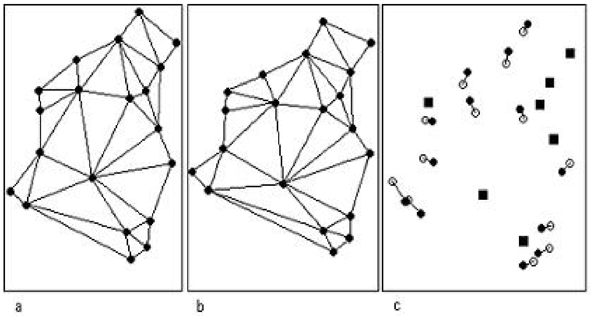


Figure 2. a – minutiae of the first image, b – minutiae of the second image, c – comparison of two minutiae configuration after rigid transformation, bars – points that are closer then 2 pixels, dark circles – positions of minutiae in the first image, blank circles – in the second image

The absolutely different approach was suggested by R. Cappelli, D. Maio and D. Maltoni [15]. They developed analytical model of fingerprint deformation. But it has sufficient shortcomings, for example, irreversibility even of small deformations.

Our article is mainly devoted to the registration of deformations. We propose an algorithm of restoration of deformations knowing correspondent points in two images. As far as fingerprints have regular structure consisting of ridges and valleys, often it is virtually impossible to find more then 50 matching points. Apparently, only minutiae can be considered as matching points, all other ways of finding correspondences used in pattern recognition (points of maximal and minimal curvature etc.) are unstable in respect to elastic deformations. In examples minutiae were matched manually, during statistical analysis of deformation, algorithm of automatic finding of correspondences is applied.

2. Model of Elastic Deformation

In general the dynamics of a small elastic deformation is considered to satisfy the Navier linear elastic PDE:

$$Lu(x,y,z,t) = -f(x,y,z), \qquad (1)$$

where *L* is the following differential operator:

$$L = \mu\nabla^2 + (\lambda+\mu)\nabla\,\mathrm{div} - \rho\frac{\partial^2}{\partial t^2}, \quad (2)$$

u is the vector of displacement; *f* is the external force. Coefficients λ and μ are the Lame's elasticity constants. These parameters can be interpreted in the terms of Young's modulus *E* and Poisson's ratio ν

$$E = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu}, \qquad (3)$$

$$\nu = \frac{\lambda}{2(\lambda+\mu)}. \qquad (4)$$

In fact a fingerprint image is captured when finger is immobile, it means that the partial derivative $\rho\frac{\partial^2 u}{\partial t^2} = 0$. Such solutions are called steady state and they do not depend on time *t*, i.e. $u(x,y,z,t) = u(x,y,z)$. In this case the Navier PDE has the following form:

$$\mu\nabla^2 u + (\lambda+\mu)\nabla\,\mathrm{div}\,u + f = 0. \qquad (5)$$

Unlike plastic materials solution of equation (1) for elastic material depends only on current force distribution and does not depend on previous configurations ("history").

Investigating properties of fingerprint deformations, it is possible to neglect 3D structure of finger and to consider 2D model for area of contact of finger and scanner surface. In fact this area carries the main information available for further processing.

Obviously in the 2D model all displacements of tissue are located in the plane of contact. Such restriction is called plain strain. The different sort of 2D elastic problem is plane stress, when the material is plane and pressure is orthogonal to this plane. The plane stress restrictions are normal for studying of dynamics of metal plates and exact solution can be found using the TPS. So the TPS is the solution of problem that is absolutely different from registration of elastic deformation of soft tissues. It is one of the possible reasons why the TPS are hardly applicable to studying of fingerprint deformations.

As was mentioned above, in case of elastic material, deformation depends only on current configuration, so a fingerprint deformation can be fully described by the function of displacement:

$$f : X \to R^2 . \qquad (6)$$

Let us define the vector $(u(x,y), v(x,y))$ of displacement at the point (x, y):

$$(u,v) = f(x,y) - (x,y) . \qquad (7)$$

The strain tensor $\tilde{\varepsilon}$ is defined by the next formula:

$$\tilde{\varepsilon} = \begin{pmatrix} \frac{\partial u}{\partial x} + \frac{1}{2}\left(\frac{\partial u}{\partial x}\right)^2 & \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\right) \\ \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial x}\right) & \frac{\partial v}{\partial y} + \frac{1}{2}\left(\frac{\partial v}{\partial y}\right)^2 \end{pmatrix} . \qquad (8)$$

The linear approximation of (8) is the following tensor:

$$\varepsilon = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) \\ \frac{1}{2}\left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x}\right) & \frac{\partial v}{\partial y} \end{pmatrix} . \qquad (9)$$

Let us assume that the material reveals linear dependence between pressure and strain (what is almost true for small deformations of biological tissues). In that case the pressure tensor σ can be calculated using the following formula:

$$\sigma = \begin{pmatrix} \frac{E(1-\nu)}{(1+\nu)(1-2\nu)}\left(\varepsilon_1^1 + \varepsilon_1^2\right) & \frac{E}{2(1+\nu)}\varepsilon_1^2 \\ \frac{E}{2(1+\nu)}\varepsilon_2^1 & \frac{E(1-\nu)}{(1+\nu)(1-2\nu)}\left(\varepsilon_2^2 + \varepsilon_2^1\right) \end{pmatrix} . \qquad (10)$$

Vector of involved forces is

$$F = \begin{pmatrix} f_x \\ f_y \end{pmatrix} . \qquad (11)$$

The overall energy E_0 and energy E_d of deformation are determined by the following formula:

$$E_0 = -A + E_d = -\int_S (uf_x + vf_y)dS + \\ + \frac{1}{2}\int_S (\varepsilon_1^1\sigma_1^1 + \varepsilon_2^2\sigma_2^2 + \varepsilon_1^2\sigma_1^2)dS. \qquad (12)$$

In case of linear isotropic material the energy of deformation is homogeneous quadratic form that depends only on the strain tensor elements. Also it is natural to assume that the form is invariant with respect to orthogonal transformation.

$$E_d = \frac{1}{2}\int_S \left(c_1\left(\varepsilon_1^1\right)^2 + c_2\left(\varepsilon_2^2\right)^2 + c_3\left(\varepsilon_1^2\right)^2 \right)dS + \\ + \frac{1}{2}\int_S \left(c_4\varepsilon_1^1\varepsilon_1^2 + c_5\varepsilon_2^2\varepsilon_1^2 + c_6\varepsilon_1^1\varepsilon_2^2\right)dS. \qquad (13)$$

Apparently, the coefficients must satisfy the following conditions:

$$c_1 = c_2; \\ c_4 = c_5; \\ 4c_3 = c_1; \\ 2c_6 = c_1. \qquad (14)$$

The two independent coefficients c_1 and c_4 are determined by the internal properties of material

$$c_1 = c_2 = \frac{E(1-\nu)}{(1+\nu)(1-2\nu)}; \qquad (15)$$

$$c_4 = c_5 = \frac{E\nu}{(1+\nu)(1-2\nu)} . \qquad (16)$$

As is known [16] solution of Navier elastic PDE (5) minimizes the energy (12). There is no idea how determine operating forces in the automatic verification systems. One of the approaches is minimizing the function E_d of deformation energy. Without any additional constrains the function E_d is minimized by zero function of displacement. In our case additional restrictions are correspondent points of two images:

$$p_i + (u(p_i), v(p_i)) = q_i ,$$

where $\{p_i\}$ is the set of points in the first image and $\{q_i\}$ is the correspondent set in the second image.

Let us consider the following functional that reflects deformation:

$$W(u,v) = E_d(u,v) + \lambda\, S(u,v) , \qquad (17)$$

where E_d is deformation energy and

$$S(u,v) = \sum_{i=1}^{n}\left(p_i + (u(p_i), v(p_i)) - q_i\right)^2 \qquad (18)$$

reflects the measure of approximation. Coefficient λ shows the importance of the approximation component.

The minimum of W can be found numerically using finite elements method (FEM). The displacement is defined on rectangular lattice and interpolated to the entire image with for example bilinear splines.

3. Implementation

The elastic model of fingerprint deformations is applied to the pair of deformed fingerprints (with manual positioning of correspondent minutiae) and three sets of fingerprints:

1. Subset of the BioLink Database (100 sets of 3 images of each fingerprint)
2. FVC2002 DB1 (100 sets of 8 images of each fingerprint) [17],[18]
3. Set of strongly deformed fingerprints (123 images of 10 fingerprints)

The correspondent points of images in all three databases are found using BioLink algorithm [19].

Young's modulus and Poisson's ratio for human skin have some variations. Young's (or *E*-) modulus depends on age and usually changes from 6 to 11 kg/mm^2. As is clear from (9), (10) and (13) E does change the solution of equation, it changes only absolute value of energy. In fact, the value of Young's modulus can be included into coefficient λ. Poisson's ratio ν for human skin is considered to vary approximately around 0,33. For the purposes of numerical calculation the mean values are taken, E=9[kg/mm^2], ν=0,33. The input images are processed to the 300x400 size (500dpi) and are divided into 120 elements of size 10x10.

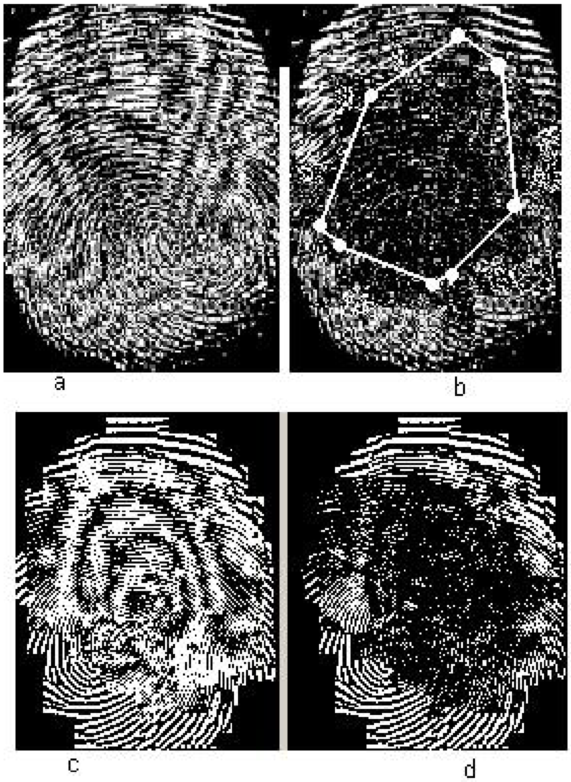


Figure 3. Direct overlapping of two images within the region of minutiae correspondence: a – without registration of deformation; b – with registration of deformation; c – filtered without registration of deformation; d – filtered with registration of deformation.

The binary correlation can be used as measure of identity of two deformed images. In the figure 3 picture a shows the direct overlapping of images after rigid transformation, picture b – after registration of elastic deformation. The results show that binary correlation of two images sufficiently increases inside the convex hull of correspondent points. At the same time after registration of deformation correspondent minutiae of two images become virtually congruent.

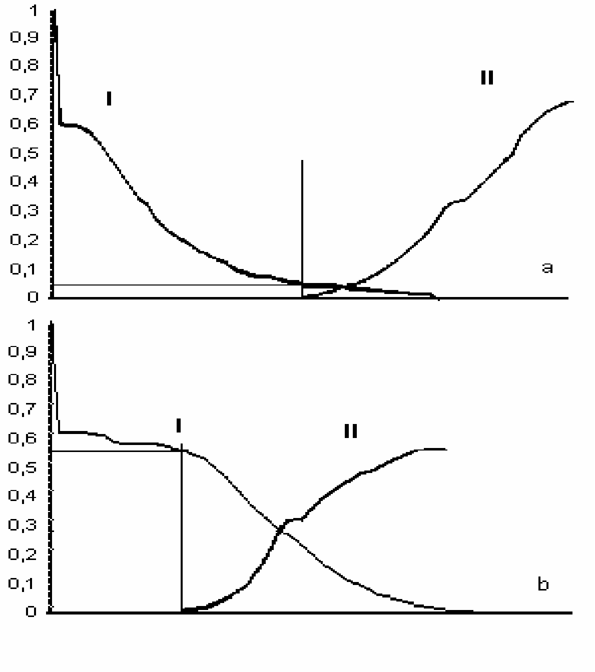


Figure 4. BioLink Database. a – distributions of energy of deformation, b – distributions of overall energy, I – entire database, II – strongly deformed images.

During automatic analysis of large sets of deformed images the binary correlation is less suitable for performance evaluation then in case of the manual demonstration, because it suffers from the following factors:

1. Different quality of images caused by different conditions of application (temperature, humidity etc.).
2. Different input devices.
3. Minutiae extraction precision. If correspondent points are determined with 2-3 pixels precision, the algorithms that compare minutiae structure work well. At the same time, binary correlation of entire image can sufficiently decrease because of improper estimation of both the rigid transformation and the deformation.

As far as the main task of the current work is evaluation of the measure of deformations appearing in the real applications, integral characteristics (overall energy and deformation energy) are calculated. These values are calculated for genuine matches of Database1 (BioLink's one) and Database3 (set of strongly deformed images). The distributions of energies are represented in the figure 4. As is clear from charts in real applications only small share (about 4 percent) of fingerprints have real deformation. Ignoring these applications it is possible to use algorithms with extremely low FAR.

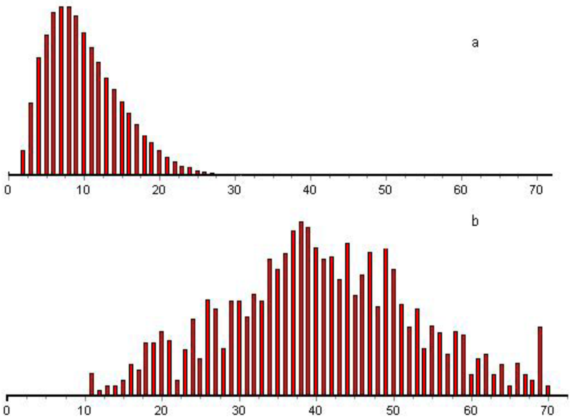


Figure 5. Number of correspondent minutiae, a – genuine matches, b – impostor matches.

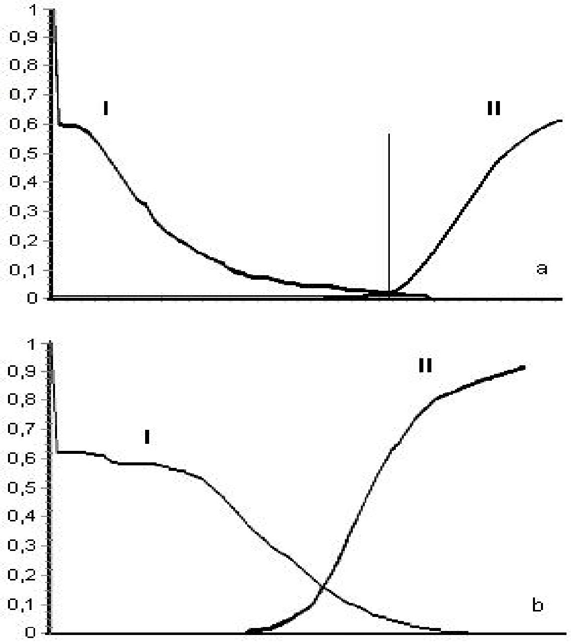


Figure 6. BioLink Database. a – distribution of deformation energy, b – distribution of overall energy, I – genuine matches, II – impostor matches.

At the same time formal evaluation of deformation can be used as the measure of similarity of minutiae configurations. In genuine matches it is more or less clear how correspondent points can be found. The same techniques are applicable to impostor matches as well. In the situation when the storing of entire image (or even parts) in database is prohibited what is natural for AFAS, analysis of minutiae configurations may be used.

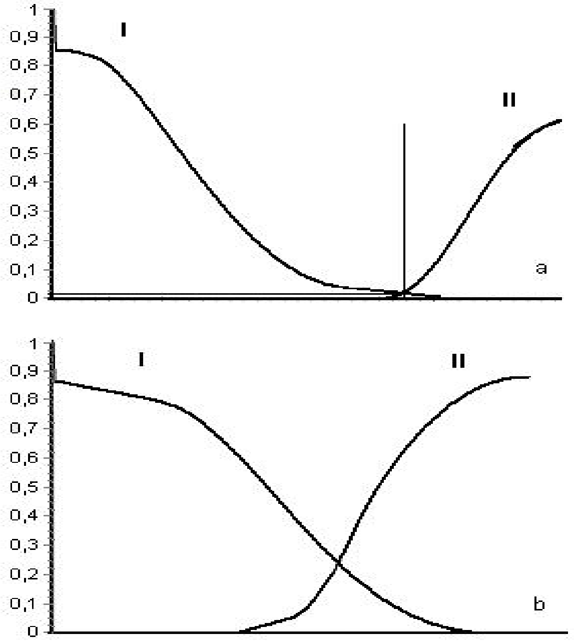


Figure 7. FVC2002 DB1. a – distribution of deformation energy, b – distribution of overall energy, I – genuine matches, II – impostor matches.

In this case the template might store the coordinates of points, number of characteristics of points (such as direction angle), distances between some points (measured for examples in ridges). If quality of input fingerprints is low due to hardware imperfectness many of parameters mentioned above are unstable. Even some points cannot be detected.

Deformation energy can be used as one of the measure of correspondence of minutiae structures. Formally, values of overall and deformation energies can be calculated even for impostor application if we have two sets of points which seem to be correspondent. Besides energies the important output is the number of correspondent minutiae (figure 5). Distributions of energies are represented in figures 6 (BioLink Database) and 7 (FVC2002 DB1). From FVC Database some genuine applications (about 15%) were removed because of small area of intersection (less then 4 minutiae in intersection). If intersection of two images is small there is no sense in studying deformations because small parts of images usually have relatively small deformations.

As is clear from figures, overall energy is much less informative parameter then pure energy of deformation. In fact the value of deformation energy of minutiae configurations can be used as auxiliary score parameter in matching algorithms. Sole energy of deformation provides Equal Error Rate (EER) equal approximately to 1% on BioLink Database (more precise EER can be confidently defined because of small number of tests). On FVC2002 DB1 EER is about 1,5%. Removing of 15% of genuine matches can only lower EER because these matches certainly have deformation energy much less then threshold. The values of energy for those matches were not calculated because it demands sufficient modification of procedure to the case of small number of correspondent points.

4. Conclusion

The conducted testing of the model of elastic deformations shows good correspondence to the real fingerprint deformations.

In natural input conditions only relatively small share of fingerprints are deformed (4%, Figure 4). In this case deformations do not lower system performance sufficiently. It means that bounds of possible deformations are not unlimited and can be calculated precisely. And, therefore, the complete deformational invariance is not a necessary condition for the most of biometric applications.

The energy of deformation can be used as auxiliary score in matching algorithms. It might bring quite good performance improvement. On the other hand, the proposed model and the estimation for p.d.f. of deformation energy allow to predict the performance of any matching algorithm.

5. References

[1] Lee H.C. and Gaenssley R.E., *Advances in Fingerprint Technology*, Elsevier, New York, 1991.

[2] Halici U., Jain L.C., Erol A., *Introduction to Fingerprint Recognition, Intelligent Biometric Techniques in Fingerprint and Face Recognition*, CRC Press, 1999.

[3] Eleccion M., "Automatic Fingerprint Identification", *IEEE Spectrum*, 1973, 10, pp. 36-45.

[4] Jain A.K., Hong L. and Bolle R., "On-Line Fingerprint Verification", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 1997, 19(4), pp. 302-314.

[5] Jain A.K., Hong L., Pankanti S. and Bolle R., "An Identity-Authentication System Using Fingerprints", *Proc. of IEEE*, 1997, 85(9), pp. 1365-1388.

[6] S. Pankanti, S. Prabhakar and A.K. Jain, "On the Individuality of Fingerprints", *IEEE Trans. PAMI*, 2002, 24(8), pp. 1010-1025.

[7] Daugman J., "The Importance of Being Random", *Pattern Recognition*, vol.36, no.2, 2003.

[8] *Large & Medium Scale ID*, online available at: www.biolinkUSA.com.

[9] *Civil Fingerprint Identification Systems*, online available at: www.east-shore.com.

[10] www.civilidsystems.com.

[11] Bazen A.M., Gerez S.H., "Thin-Plate Spline Modelling of Elastic Deformation in Fingerprints", *Proceedings of 3rd IEEE Benelux Signal Processing Symposium*, 2002.

[12] Bookstein F.L., "Comment to D.G. Kendall's A survey of the statistical theory of shape", *Statistical Science*, vol.4, no. 2, 1989, pp. 99-105.

[13] M. Fornefett, K. Rohr and H.S. Stiehl, "Elastic Medical Image Registration Using Surface Landmarks with Automatic Finding of Correspondences", In A. Horsch and T. Lehmann, editors *Proc. Workshop Bildverarbeitung fur die Medizinl, Informatik actuell, Munchen, Germany*, Springer-Verlag Berlin Heidelberg, 2000, pp. 48-52.

[14] M. Fornefett, K. Rohr and H.S. Stiehl, "Radial Basis Functions with Compact Support for Elastic Registration of Medical Images", *Image and Vision Computing*, 19 (1-2), 2001, pp. 87-96.

[15] Raffaele Cappelli, Dario Maio, Davide Maltoni, "Modelling Plastic Distortion in Fingerprint Images", *ICAPR2001*, pp. 369-376.

[16] Shames, I.H. and Pitarresi, J.M., *Introduction to Solid Mechanics*, Upper Saddle River, NJ, 2000.

[17] First International Competition for Fingerprint Verification Algorithms (FVC2000), bias.csr.unibo.it/fvc2000/.

[18] FVC2002, the Second International Competition for Fingerprint Verification Algorithms (FVC2000), bias.csr.unibo.it/fvc2002/.

[19] U.S. Patent #6,282,304

A Study on the Fusion of Sequential Fingerprints Enrolled with Rolling and Sliding

Hunjae Park, Kyoungtaek Choi, Sanghoon Lee and Jaihie Kim
Department of Electrical and Electronic Engineering , Yonsei University
Biometrics Engineering Research Center, Seoul, Korea
hunjae_park@daum.net

Abstract

Fingerprint-based verification systems are used commonly in the field of biometrics. Especially, a small-sized sensor makes possible use in embedded systems, but it does not provide sufficient information for high accuracy user verification. In this paper, to obtain a wide region of fingerprint, the new fingerprint enrolling scheme which includes rolling method as well as sliding on a small-sized sensor is proposed and a block matching algorithm for estimating an alignment parameter is also proposed. After aligning the two images, the next image is warped in order to compensate for deformation using a two-pass mesh warping algorithm. Our experiments show that the number of minutiae is increased against methods to integrate multiple impressions.

Keywords – fingerprint fusion, rolling, block-matching, warping, deformation, coherence

1. Introduction

Fingerprint-based verification systems are provided widely since they are convenient to use and relatively superior to other biometrics systems in terms of price and performance. Especially, a small-sized sensor(e.g., solid-state sensors) has the advantage that it can be used in many application fields(e.g., laptops, cellular phones). However, a limited amount of information about the fingerprint is available due to the small physical size of the sensing area (Fig.1). Therefore the relatively amount of small overlap between the template and query impressions results in degraded performance, like a higher rate of false rejects and/or false accepts. To overcome this problem, some researchers proposed algorithms to integrate multiple impressions from the same finger, but this method has little effect on impressions sensed from the similar portion of a specific finger. Furthermore, it is very hard to integrate impressions sensed from very different portions of a specific finger.

To overcome the problem, we use a method in which the user rolls his or her finger on the sensing area (the axis of rolling is not moved) to obtain images sensed form which subsequently diffused portion of the fingerprint used for enrollment (Fig.2). That is to say, the user rolls his or her finger from one edge to the other edge, simultaneously. A key factor is that this sliding must not exceed the boundaries of a small-sized sensor area. We can obtain a sequence of partial fingerprint images using this method and attain one template image from fusing them.

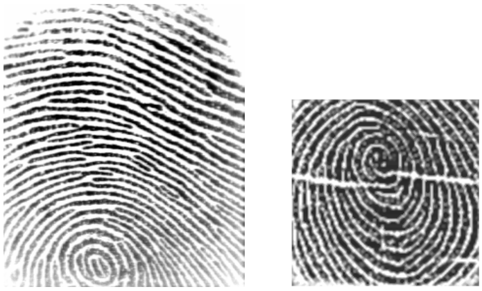

(a) (b)

Figure 1. Fingerprint images obtained from sensors that have large sensing area and small sensing area: (a) large fingerprint image. (b) small fingerprint images.

This enrollment method using rolling included with sliding has the following advantages:

(a) It can obtain a wide area of fingerprint more stably than existing general fusing method from multiple impressions.

(b) Two temporally adjacent images are highly correlated which makes easy to align the images using their correlation ratio.

The conventional fusion methods used currently in our field of study are summarized as follows: Jain et al. proposed an alignment algorithm using ICP to construct a composite fingerprint template while using multiple impressions[1]. Ramsor et al. proposed an alignment algorithm using the RANSAC method and a combination method of the minutiae information[2]. Qun et al. proposed an alignment algorithm using a Clique Graph and an information fusion method, utilizing a clustering algorithm[3]. Lee et al. proposed an alignment algorithm using a Distance Map derived form ridge information [4].

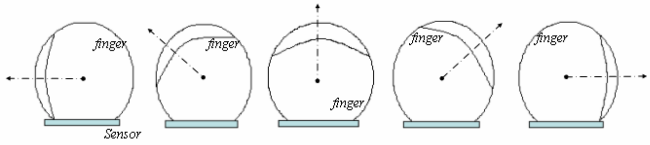


Figure 2. The finger rolls on the sensing area without moving the rolling axis.

Above fusion methods are entirely based on minutiae when aligning two images. Despite of large overlapped images, however, the minutiae based-alignment algorithm may compromises misaligning due to the insufficient corresponding minutiae pairs. Local deformation (e.g., by the movement) will also result in erroneous error without compensating for deformation. In this paper, sequential fingerprints enrolled with rolling and sliding are aligned using a block matching scheme instead of minutiae-based matching. Local deformation is compensated and then minutiae are extracted. In this case, the misalignment by insufficient minutiae does not occur. Furthermore, it is more time efficient method since it does not extract feature from each impression.

Our paper is organized as follows. In Section 2, we describe the alignment procedure using block matching scheme. In Section 3, we describe the deformation compensation procedure using image warping. In Section 4, we describe the image fusion process. The experimental results are shown in Section 5. Finally, Section 6 contains our conclusions.

2. Alignment

A fingerprint being translated and rotated changes its appearance while sequential images are captured. Therefore the process that aligns coordinate systems of impressions is absolutely necessary in order to integrate the sequential fingerprint images used for enrollment.

When aligning a sequence of images at the enrollment step, we can utilize both minutiae and intensity image while we can only use minutiae at matching step. Using additional intensity information of fingerprints can increase aligning accuracy. An exact alignment is very important since to fuse sequential fingerprints is a significant part of procedure that template for matching is made and an inexact template can cause a higher false reject rate. In this paper, sequential images of a fingerprint are aligned at the raw data level by using a block matching algorithm.

2.1. Block Matching

Each pair of adjacent images in sequential fingerprints has a high mutual similarity because of a short difference in time, and this similarity enables us to utilize the correlation between them at the alignment step.

To obtain a correlation between two fingerprints based on the entire image is quite a time-consuming computation process. For this reason, it is preferable to divide the image into blocks and prosecute block matching to estimate the most adequate alignment parameters. These blocks are sensitive to local deformation, but this sensitivity of block can be applied to compensate for the local deformation of fingerprints.

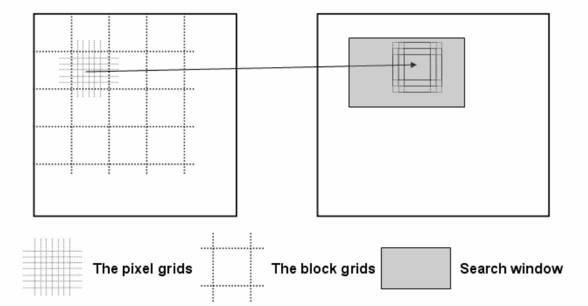


Figure 3. Two-dimensional illustration of the block matching scheme.

When two images *P* and *Q* are aligned, respectively, *P* is divided into blocks. Each block of P matches image *Q* in pixel-wised block. That is, each block is translated and rotated into several positions within the searching window, and then compared to the corresponding block of *Q* (Fig. 3). Normalized cross-correlation can be used as a similarity measure. At each position, an intensity correlation coefficient score, *CC*, between the block and the corresponding block is computed as [5]:

$$CC = \frac{\sum_x (I_P(x,y) - \mu_P)(I_Q(x,y) - \mu_Q)}{\sqrt{\sum_x (I_P(x,y) - \mu_Q)^2}\sqrt{\sum_x (I_P(x,y) - \mu_Q)^2}} \qquad (1)$$

As mentioned earlier, each pair of adjacent images has only small translation and, especially, rotation, which make it possible to apply a full search within the available searching area. In order to determine coefficients of a alignment transform *Q* from *P*, each block computes *CC* scores for all available transformations. If *CC* score is higher than a certain threshold, we can consider such parameters as the candidate coefficient of a true transform.

Each block has locally small deformation thanks to the small amounts of translation and rotation. In this case, it makes sense to assume that the all candidates are distributed in the given translation domain and this distribution is centered at a global transform. Consequently, the pdf(probability density function) of a transform parameter can be gained using the Parzen density estimation. It is peaked at true parameters of an alignment transform.

2.2. Parzen Windows

The formula used to estimate pdf is as following [6]:

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) \tag{2}$$

where, n is the number of samples and $\varphi(\mathbf{u})$ is the kernel function of the Parzen window. The symbol h_n is the parameter concerning window width. Eq.(2) can be modified to make it suitable for use in this paper. It follows that:

$$p_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{m_i}\frac{1}{h_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_{i,k}}{h_N}\right) \tag{3}$$

where, $N(=\sum_{i=1}^{n} m_i)$ are the number of whole candidates and candidates of the i^th block, and n is number of block and $\mathbf{x}$ is a transform parameter. Subsequently, $\mathbf{x}_{i,k}$ is a transform vector of the k^th candidate at the i^th block. It is available for weight correlation of the window function since a higher correlation ratio means a higher probability when that candidate is equal to the global transform. It follows that:

$$p_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{M_i}\left(CC_{i,k}\times\frac{1}{h_N}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_{i,k}}{h_N}\right)\right) \tag{4}$$

where, $CC_{i,k}$ is the correlation value of the k^th candidate at the i^th block. The symbol N and h_N can be negligible because it has the same value against all $\mathbf{x}$. The global transform consists of the parameter which shows maximum probability.

$$\mathbf{x}_{global} = \arg\max_{\mathbf{x}}\left(\sum_{i=1}^{n}\sum_{k=1}^{M_i}\left(CC_{i,k}\times\varphi\left(\frac{\mathbf{x}-\mathbf{x}_{i,k}}{h_N}\right)\right)\right) \tag{5}$$

where, $\mathbf{x}_{global}$ is the global alignment parameter.

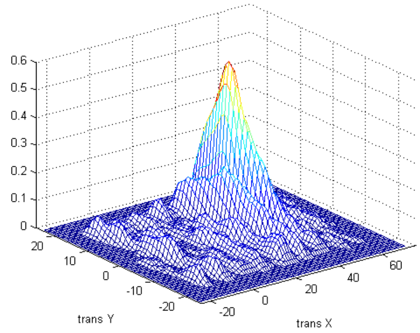


Figure 4. PDF of transform parameter.

3. Compensation for local deformation

Since the alignment transform is not a simple linear transform, one global transform does not represent the local deformation. Therefore, to compensate for local deformation is followed after the global alignment is completed.

3.1. Regularization of block transforms

It is possible to regard a local deformation as the difference between the global transform and a block transform. Therefore it is necessary to identify individual block transforms. However, for each block, the maximum similarity score does not necessarily correspond to the best transform, partly because of noise and the deformation in each block. A regularization step is necessary and this is achieved by taking into account the influence of the transform of the neighbor blocks. The Parzen density estimation can be used for a regularization method. Our paper also uses a hierarchical approach to identify the local block transform (Fig.5).

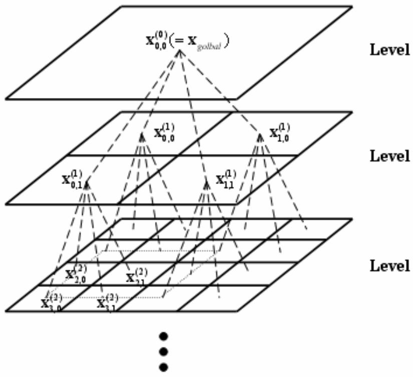


Figure 5. Hierarchical structure in order to identify the local block transform.

First of all, all blocks are divided four partitions. And then, transform parameter $\mathbf{x}_{i,j}^{(n)}$ of each partition can be estimated using Parzen window with candidates of blocks included or adjoined the partition. This transform should be close to the transform $\mathbf{x}_{\lfloor i/2\rfloor,\lfloor j/2\rfloor}^{(n-1)}$ at higher level. Hence, the parameter having maximum probability in near at the partition's transform at higher level regars as the partition's transform at current level. Such process repeats until the number of block included partition is less than or equal to four. Each block transform is finally estimated with its candidates and eight neighbor blocks' candidates.

3.2. Image warping

After block transforms are estimated, local deformations are compensated using point-based warping technique. Every center points of each block are considered as corresponding points when warp image *Q*', that is, points derived from global transform are utilized as destination points while source points are defined as translated points by block transforms (Fig.6).

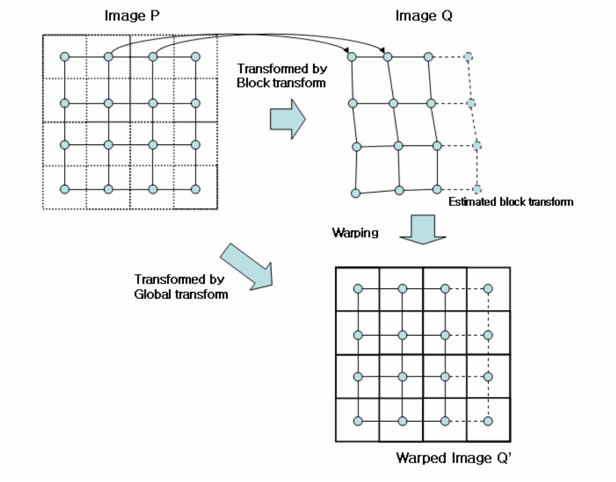


Image P
Image Q
Transformed by Block transform
Estimated block transform
Warping
Transformed by Global transform
Warped Image Q'

Figure 6. Illustration of the warping scheme.

This paper use two-pass mesh warping based on an algorithm represented in [7]. This algorithm uses Fant's resampling algorithm and cublic spline as the interpolation method.

4. Image fusion

After the subsequent fingerprint images are aligned as described in section 2 and compensated for as explained in section 3, the warped image *Q*' is transformed into the coordinate system of P by global transform. An intensity value at each pixel in the fusion image is computed using the weighted sum of corresponding pair. Since coherence of ridge orientation can reflect the quality of an individual image, it is possible to use of coherence as weights [8].

$$I_F(x,y) = \frac{Coh_P(x,y)I_P(x,y) + Coh_{Q'}(x,y)I_{Q'}(x,y)}{Coh_P(x,y) + Coh_{Q'}(x,y)} \quad (6)$$

where $I_F(x,y)$, $I_P(x,y)$, and $I_{Q'}(x,y)$ is intensity values at pixel of fusion image, image *P*, warped image *Q'* in (x,y). And coherence is compute as :

$$Coh = \frac{\sqrt{(G_{xx} - G_{yy})^2 + 4G_{xy}^2}}{G_{xx} + G_{yy}} \quad (7)$$

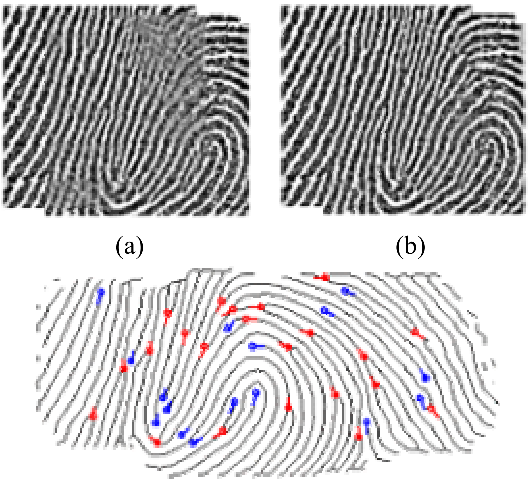
$$\begin{bmatrix} G_{xx} & G_{yx} \\ G_{xy} & G_{yy} \end{bmatrix} = \sum_W \begin{bmatrix} G_x^2 & G_x G_y \\ G_x G_y & G_y^2 \end{bmatrix} \quad (8)$$

where, G_x and G_y is x and y element of the gradient vector in the Cartesian coordinate. And the symbol *W* is the window size.

Finally, a single template image is obtained by re-integrating the fused images.

5. Experimental results

A set of fingerprint images using rolling scheme is acquired through solid-state fingerprint sensor manufactured by AuthenTec. This sensor is acquired about 6.5 images per second. The size of the image is 192x192 pixels with the resolution of 500 dpi and 72 rolling sequences are used in experiment. All rolling sequences are consisted of 2095 fingerprint images and each sequence is consisted of avg. 29.1 fingerprint images. This paper uses time-sampled 7 images from each sequence since integrating many images make fusion image blur.



(a) (b)

(c)

Figure 7. Sample fused images :

(a) two images without compensating for deformation
(b) two images with compensating for deformation
(c) sequential images

The proposed paper uses image normalization as preprocessing before the alignment. In the alignment, the 16x16 sized blocks is used to find candidates and alignment transform is calculated using Gaussian Parzen window which is sized 3. The transform is formed one parameter vector, and it is insufficient to show relation between two sequent images. Therefore, this paper estimates block transforms and warp the second image using them. In this stage, this paper uses 3-level hierarchical regularization method and 2-pass mesh warping algorithm. Finally, two images are fused one using the coherence which is calculated in the 16x16 sized windows. A sample fused image is shown in Figure 7.

The following table lists a few statistics about the fusion image generated using the block matching scheme. The number of minutiae per fingerprint increases from 14.1 for one impression to 32.8 for fused fingerprints.

Table 1. Result of the fusion

	Avg. no. of minutiae
Impression	14.1
Composite image from a sequence	32.8
Composite image from 7 impressions	30.0

6. Conclusion

We have described a new enrollment scheme using rolling in fingerprint-based verification system and a nonminutiae-based fusion method for the sequential fingerprint images. Experimental results show that this method extract more minutiae than methods to integrate the multiple impressions. It means a wider area of finger is obtained by rolling than by other methods and the amount of overlap between the template and query impressions increase.Future work involves studying the coarse alignment before the alignment step in order to decrease of processing time. The coarse alignment makes possible a small search window for block matching. We are also attempting to normalize the ridge density at a final fused image since rolling scheme makes ridge density change because of the friction between a finger and a sensor.

7. Acknowledgements

This work was supported (in part) by Biometrics Engineering Research Center,(KOSEF).

8. References

[1] A.K. Jain and A. Ross, "Fingerprint Mosaiking", *Proc. International Conference on Acoustic Speech and Signal Processing(ICASSP),* vol. 4, pp.4064-4067, 2002

[2] H. Ramoser, B. Wachmann, H. Bischof, "Efficient Alignment of Fingerprint Images", *Proc. 16th International Conference on Pattern Recognition,* vol. 3, pp. 748-751, 2002

[3] R. Qun, T. Jie, H. Yuliang and C. Jiangang, "Automatic Fingerprint Identification Using Cluster Algorithm", *Proc. 16th International Conference on Pattern Recognition,* vol. 2, pp. 398-401, 2002

[4] Dongjae Lee, Sanghoon Lee, Kyoungtaek Choi and Jaihie Kim, "Fingerprint Fusion Based on Minutiae and Ridge for Enrollment*", LNCS on Audio-and Video-Based Biometric Person Authentication,* vol.2688, pp.478-485, Jun. 2003

[5] Guofang Xiao, J. Michael Brady and J. Alison Noble "Nonrigid Registration of 3-D Free-hand Ultrasound Images of the Breast" *IEEE Transactions on Medical Imaging,* vol. 21, no. 4, pp. 405-412 April 2002

[6] Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern Classification,* Wiley-Interscience, 2002

[7] George Wolberg, *Digital Image Warpin*g, IEEE Computer Society Press, 1988

[8] Asker M. Bazen and Sabih H. Gerez "Systematic Methods for the Computation of the Directional Fields and Singular Points of Fingerprints", *IEEE Transations on Pattern Analysis and Machine Intelligence,* vol. 24. no.7, pp.905-919, July 2002

On-Line Signature Verification Method Based on Discrete Wavelet Transform and Adaptive Signal Processing

Isao Nakanishi
Faculty of Education and Regional Sciences,
Tottori University, 4-101 Koyama-minami,
Tottori-shi, 680-8551 Japan

Naoto Nishiguchi, Yoshio Itoh, Yutaka Fukui
Faculty of Engineering,
Tottori University, 4-101 Koyama-minami,
Tottori-shi, 680-8552 Japan

Abstract

This paper presents the on-line signature verification method based on the Discrete Wavelet Transform (DWT) and the adaptive signal processing. Time-varying pen-position signals of the on-line signature are decomposed into sub-band signals by using the DWT. Individual features are extracted as high frequency signals in sub-bands. This makes difference between a genuine signature and its forgery remarkable. However, there is fluctuation in number of strokes even in the genuine signature. In this paper, we introduce the Dynamic Programming (DP) matching method to suppress the fluctuation. Verification is achieved by whether the adaptive weight converges on one. However, its convergence characteristics depend on the step size parameter of the adaptive algorithm. Therefore, the normalized step size parameter by the signal power of the input signature is introduced to guarantee the convergence. Results of verification show that the verification rate of 95% is accomplished even though a writer is not permitted to refer to his/her own signature and a forgery can trace the genuine signature.

1. Introduction

As the information service over internet such as the Electronic Commerce and the Electronic Data Interchange come to be used widely, the user authentication technology becomes quite important. Until now, the memory such as password, and the belongings including a key and a magnetic card have been used for the user authentication. However, they have danger of losing and forgetting. Thus, the biometrics has attracted attention [1].

The biometrics is divided roughly into two types. The fingerprint, the iris and the retina are included in static biometrics. They achieve high verification rate while they require special detective devices. Therefore, the use of them is limited to the financial institution or the special facilities where secret information is defended. The voice-print and the signature are of dynamic biometrics. The user authentication by the voice-print is effective especially on the service with a telephone but it requires counter measures to such problems as recorded voice and surrounding noise.

The user authentication by the signature consists of two types. The off-line type has been researched as a target of the pattern matching, in which the shapes of written signature are compared. On the other hand, the on-line type classifies the signature by such time-varying signals as the pen-position, the pen-pressure, the pen-inclination and so on [2-5]. These contain more individual features as habits than the off-line type. Especially, the imitation of the pen-pressure or the pen-inclination is difficult for others while the pen-position can be easily traced if the shape of the signature is known. In addition, the electronic pen-tablet which is used to digitize the on-line parameters is a standard input device of the computer; therefore, the on-line type is suitable for the user authentication in the service on computer networks.

In this paper, we authenticate the user by only the pen-position parameter which is utilized for the hand-written input or the pointing even in the Personal Digital Assistants (PDA). However, if the signature is traced by a forger, the difference between a genuine signature and its forgery is not clear in the time-domain signal of the pen-position parameter. We have proposed to decompose such the time-varying signal into sub-band signals by the discrete wavelet transform (DWT) [6]. Moreover, we proposed to introduce the adaptive signal processing into the verification of signatures. In the adaptive signal processing, an adaptive weight is updated to reduce an error between an input signal and a desired one [7]. If the input signal is close to the desired one, the error becomes small and then the adaptive weight is sure to converge on one. Therefore, when both the signals are of the genuine signature, the adaptive weight is expected to converge on one. By using the convergence of the adaptive weight, we can verify whether an input signature is genuine or forged. In addition, even in genuine signatures, there is fluctuation in the number of strokes and then it degrades the performance of verification; therefore, we introduce the robust stroke matching by using the DP matching method into the verification.

This paper is organized as follows. In Sec.2, we explain to extract time-varying signals of the on-line signature, especially the pen-position parameter and make it clear that discriminating between the genuine signature and the forgery is difficult in the time-domain signals. In Sec.3, we introduce the sub-band decomposition by the DWT and show that differences between the genuine signature and the forgery become clear in the sub-band signal. In Sec.4, we explain the verification method based on the adaptive signal processing and the robust stroke matching by using the DP matching method. In Sec.5, the effectiveness of the proposed system is examined through experimental results. Finally, Sec.6 presents conclusions and future works.

2. On-line Signature

2.1 Extraction of signature parameter

An on-line signature is digitized with a pen-tablet. The specification of the pen-tablet used in this paper is presented in Table 1.

Table 1. Specification of pen-tablet

Model	WACOM Cintiq C-1500X
Method	Electromagnetic Induction
Active Area	304.1×228.1 mm
Resolution	0.05 mm
Accuracy	±0.5 mm
Report Rate	185 point/sec
Reading Height	5 mm
Pen Pressure	512 levels

Figure 1 shows the definition of the on-line parameters. In this pen-tablet, parameters of the pen-position and the pen-pressure are obtained as discrete time-varying signals. Especially, the pen-position parameter is at least provided in portable devices such as the PDA for the handwriting input and the pointing. In this paper, we identify signatures by using only the pen-position parameter.

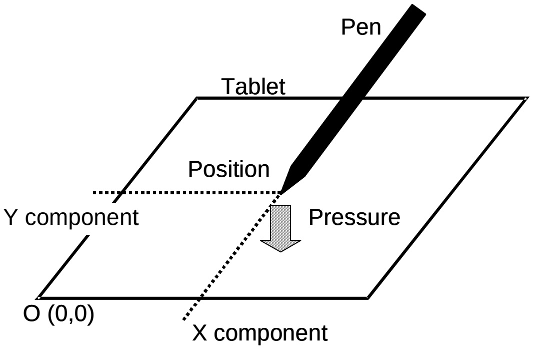


Figure 1. On-line signature parameters

Actually, the pen-position parameter consists of discrete time-varying signals $x^*(n')$ and $y^*(n')$ of x and y components, respectively. n' (= 0~$N_{\max}$) is a sampled time index and $N_{\max}$ shows its maximum value. Let us consider variations of the parameter. Even if the same person signs twice, quite the same signature parameter is not obtained because the signature is dynamic biometrics. Different writing time results in different number of data. To suppress the influence of the variation, each sampled time n' is normalized by $N_{\max}$.

$$n = n' / N_{\max} \qquad (1)$$

where n (=0~1) is a normalized sampled time index, which has real number value.

Next, in order to reduce differences of writing place and size of the signature, it is also required to normalize the amplitudes of $x^*(n)$ and $y^*(n)$. As a result, we define the normalized pen-position parameters as

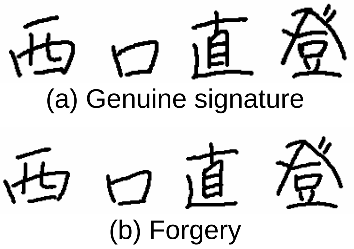
$$x(n) = \frac{x^*(n) - x_{\min}}{x_{\max} - x_{\min}} \cdot \alpha_x \quad \left(x_{\min} \le x^*(n) \le x_{\max}\right) \qquad (2)$$

$$y(n) = \frac{y^*(n) - y_{\min}}{y_{\max} - y_{\min}} \cdot \alpha_y \quad \left(y_{\min} \le y^*(n) \le y_{\max}\right) \qquad (3)$$

where subscripts max and min indicate maximum and minimum values of parameters, respectively. α_x and α_y are scaling factors for expansion of the dynamic range. The normalized signal is decomposed into sub-band signals by the DWT and then the signal power is distributed to each band. For avoiding the under flow in calculation at each band, the dynamic range of the normalized signal must be spread in advance. These factors are set 100 experimentally in this paper.

2.2 Genuine signature and the forgery

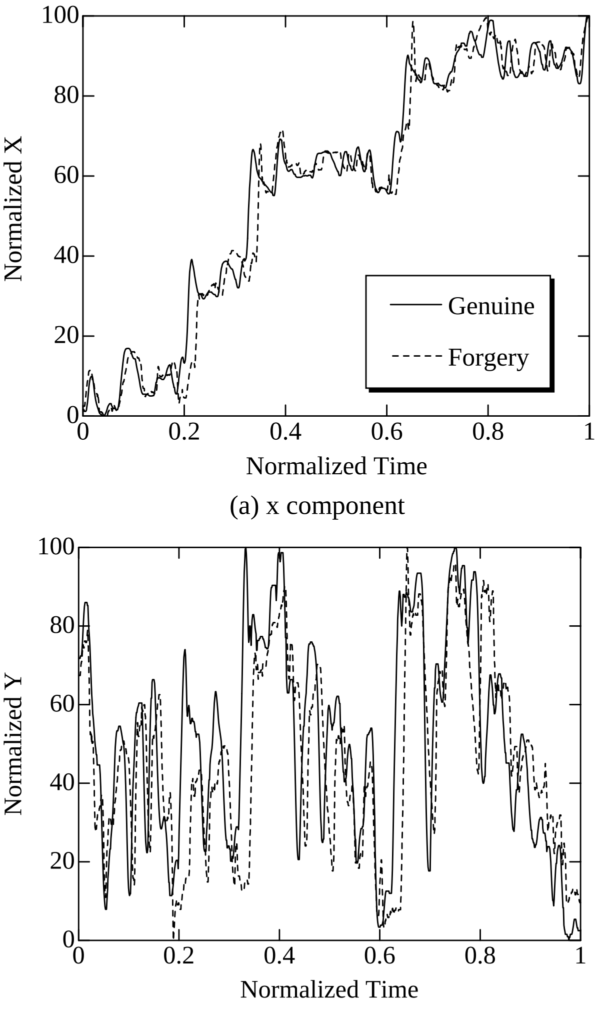
The pen-position parameter depends on the shape of the signature. Thus, it can be easily imitated if a forger traces a genuine signature. Figure 2 shows an example of them. Especially, the forgery was obtained by tracing the genuine signature. It is impossible to distinguish between the genuine signature and the forgery in comparison of signature shape.



(a) Genuine signature

(b) Forgery

Figure 2. Examples of Signatures



(a) x component

(b) y component

Figure 3. Pen-position parameters

Figure 3 shows time-varying signals of the normalized pen-position parameters $x(n)$ and $y(n)$ of the above signatures. Solid lines are of the genuine signature and dashed lines are of the forgery. These comparisons indicate that it is quite difficult to discriminate between the genuine signature and the forgery in the time-domain.

3. Enhancement of signature feature by DWT sub-band decomposition

The wavelet transform gives us a time-frequency analysis, which is effective for the non-stationary signal. In this section, we show that the sub-band decomposition by the Discrete Wavelet Transform (DWT) [6] makes difference between the genuine signature and the forgery more remarkable.

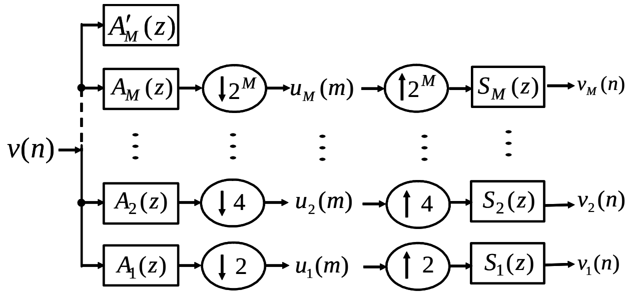


Figure 4. Parallel structure of sub-band decomposition by DWT

3.1 DWT Sub-band decomposition

In the following, $x(n)$ and $y(n)$ are represented as $v(n)$ for convenience. The DWT of the normalized pen-position component is $v(n)$ defined as

$$u_k(m) = \sum_n v(n)\overline{\psi}_{k,m}(n) \tag{4}$$

where k is a frequency (level) index. $\psi_{k,m}(n)$ is the wavelet function and $^{-}$ denotes its conjugate. It is well known that the DWT corresponds to the octave-band filter bank. The DWT pair is expressed by a parallel structure as depicted in Fig.4. ($\downarrow 2^m$) and ($\uparrow 2^m$) denote the down-sampling and the up-sampling, respectively. When $H_0(z)$ and $F_0(z)$ are transfer functions of the LPF (low pass filter), and $H_1(z)$ and $F_1(z)$ are those of the HPF (high pass filter), the synthesis filters $A_k(z)$ ($k=1\sim M$) and the analysis filters $S_k(z)$ ($k=1\sim M$) are defined as

$$A_1(z) = H_1(z) \tag{6}$$

$$A_2(z) = H_0(z)H_1(z^2) \tag{7}$$

$$A_M(z) = H_0(z)H_0(z^2)\cdots H_0(z^{2^{M-2}})H_1(z^{2^{M-1}}) \tag{8}$$

$$A'_M(z) = H_0(z)H_0(z^2)\cdots H_0(z^{2^{M-2}})H_0(z^{2^{M-1}}) \tag{9}$$

$$S_1(z) = F_1(z) \tag{10}$$

$$S_2(z) = F_0(z)F_1(z^2) \tag{11}$$

$$S_M(z) = F_0(z)F_0(z^2)\cdots F_0(z^{2^{M-2}})F_1(z^{2^{M-1}}) \tag{12}$$

At each sub-band level, an input frequency band is decomposed into a low frequency band and a high frequency one. The signal in the high frequency band is called "*Detail*" and that in the low frequency band is "*Approximation*". The octave-band decomposition is accomplished by applying such decomposition to the lower frequency band repeatedly. As a result, in M level decomposition, we obtain M *Details* $v_k(n)$ ($k=1\sim M$) in parallel as shown in Fig.4. Its frequency characteristic is described in Fig. 5.

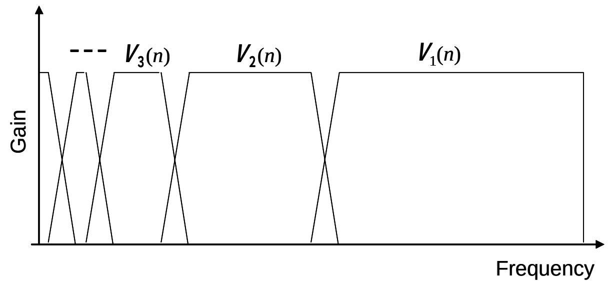


Figure 5. Octave-band decomposition

3.2 Enhancement of signature feature

The *Detail* contains differences between signals; therefore, we consider it as the enhanced feature of the on-line signature parameter. For instance, the *Detail*s at level *M* in the genuine signature and the forgery are shown in Fig.6 when each time-varying signal of x component in Fig.3 (a) is decomposed into *M* level signals by the Daubechies8 filters in the DWT. Comparing these two *Detail*s, we can confirm that the difference between the genuine signature and the forgery become remarkable by the sub-band decomposition while it is unremarkable in time-domain comparison.

The reason why the sub-band decomposition enhances the difference between signatures is as follows. In the case of forged signatures, the variation of writing time is relatively large because the writing time is not imitable. While the normalization of the writing time decreases the difference between signatures, it leads to a different sampling time in each signature. This means that an actual frequency is different from each other as shown in Fig.6 even at the same decomposition level. In genuine signatures, the variation of writing time is small, so that the actual frequency is also equivalent with each other at the same level.

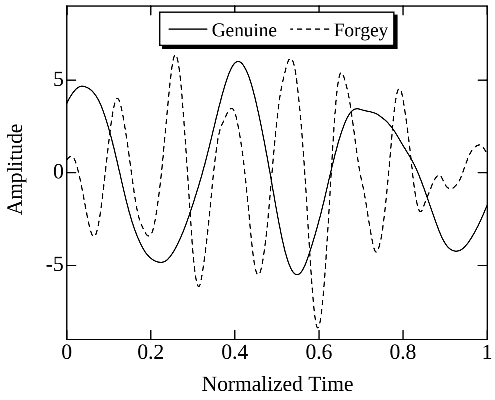


Figure 6. Example of decomposed signals

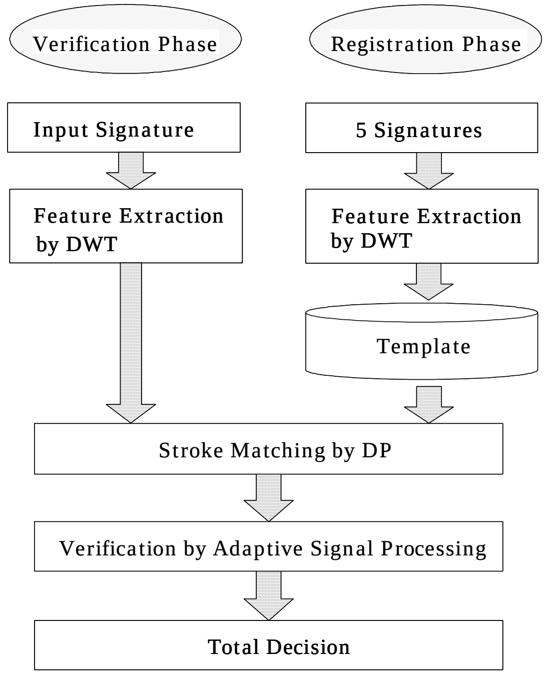


Figure 7. Flow of proposed signature verification

4. Verification method

Next, we explain about our verification method. It is unique that the adaptive signal processing [7] is adopted to discriminate between signatures at each decomposition level. In addition, the total decision of verification is achieved by considering results of x and y components at some levels, so that it is expected to be robust. Moreover, we introduce the dynamic programming (DP) matching method into the verification for robustness against variation of the number of strokes.

4.1 Flow of verification

Our signature verification method consists of two phases as shown in Fig.7. One is a registration phase and the other is a verification one. Before the verification phase, the registration phase must be accomplished. In the registration phase, five genuine signatures are decomposed into sub-band signals by the DWT. Five *Detail*s at the same level are averaged and then its results is registered as a template at each level in the data base. In the verifivation phase, an input signature is also decomposed into *Detail*s. By the way, the number of strokes is not neccesarily equal to that of the template even in the genuine signature because the on-line signature is dynamical. This causes degradation of verification performance. Therefore, the DP matching method is introduced to match the stroke of the input signature with that of the template. After stroke matching, the *Detail* of the input is compared with that of the template using the adaptive signal processing. Finally, the total decision of verification is accomplished by examining results at some decomposition levels.

4.2 Detection of stroke

For stroke matching, it is necessary to detect strokes. Each stroke consists of an intra-stroke and an inter-stroke which correspond to a pen-down condition and a pen-up one, respectively. In this paper, the stroke is detected by using the quantized pen-pressure parameter $P(n)$ as follows.

$$P(n)=\begin{cases}1 & (P^*(n)>0)\\ 0 & (P^*(n)=0)\end{cases} \tag{13}$$

where $P^*(n)$ is the pen-pressure parameter. If $P^*(n)$ is nonzero, it corresponds to the pen-down condition and so the $P(n)$ is set to 1. On the other hand, when $P^*(n)$ is zero, it indicates the pen-up condition, and $P(n)$ is also zero. Thus, each stroke can be detected as a pair of $P(n)$ =1 and $P(n)$ =0. While we utilize the pen-pressure parameter for convenience, our proposed method does not need the pen-pressure parameter essentially. The pen up/down condition can be detected by using a pen-point switch.

4.3 Making of template

Before the verification, the template must be prepared. Concretely, five genuine signatures which have equal number of strokes are decomposed into sub-band signals by the DWT. Decomposition level M is decided after examining those genuine signatures. Extracted five *Detail*s at the same level are averaged and then the result is registered as a template at each level in the data base. Incidentally, for taking an average, the number of sampled data should be equal in five signatures. However, each number of sampled data may be different from the others even in the genuine signature. To solve this problem, five signatures are averaged every intra-stroke or inter-stroke (intra/inter-stroke) in our verification system.

First, we determine the number of data in the template. Let n_i (i=1~5) be data numbers of five intra/inter-strokes, the number of data in the template n_A is given by

$$n_A=\left[\frac{1}{5}\sum_{i=1}^{5} n_i\right] \tag{14}$$

where $[x]$ is the greatest integer not greater than x.

Next, the normalized sampling period in a template stroke is given by $1/(n_A\text{-}1)$, and those in five *Detail*s are $1/(n_i\text{-}1)$ (i=1~5), too. These are illustrated in Fig.8. Five *Detail* data of which normalized sampled time $r/(n_i\text{-}1)$ (i=1~5, r=1~n_i) is the nearest to that in the template $m/(n_A\text{-}1)$ (m=0~n_A-1) are selected as described by arrows in the figure and averaged every normalized sampled time in the template. As a result, we obtain n_A data in each intra/inter-strokes. By applying this operation to all intra/inter-strokes at all level, all template data are obtained and they are registered in the database.

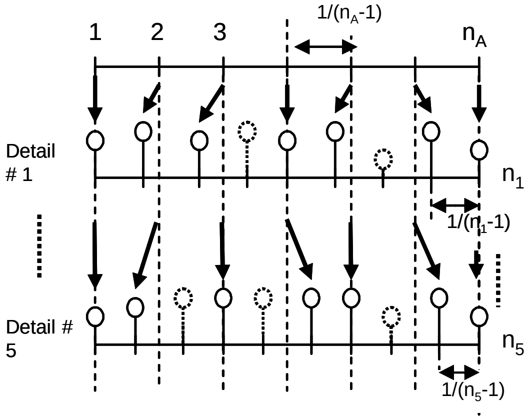


Figure 8. Making of Template

4.4 Determination of decomposition level

Next, we discuss the verification phase. The DWT corresponds to the octave-band decomposition; therefore, the decomposable level depends on the number of sampled data in an input signature. In this paper, we determine the decomposition level M of the input signature based on that in the template as given by

$$2^{M+1}\le N<2^{M+2} \tag{15}$$

where N is total number of the template data.

4.5 Stroke matching

If the number of strokes in an input signature is different from that in a template, the input signature should be considered as a forgery. However, not all genuine signatures have the same number of strokes. In fact, we confirmed that there was the stroke difference within ±2 even in the genuine signature through some experimences. Immidiately rejection of the input signature with different number of strokes causes degradation of verification performance. In this paper, we allow the input signature with the stroke difference within ±2. However, our verification is done every intra/inter-stroke and so the number of strokes in the input signature should be equal to that in the template. Therefore, we adopt the dynamic programming (DP) matching method to identify the number of strokes in the input signature and the template.

The difference number of strokes γ between the input signature and the template is calculated by using the pen up/down information $P(n)$. If $0<|\gamma|<2$, either of more strokes is decreased by coupling strokes. Figure 9 shows a case of γ=1. Assuming the number of strokes in Ψ is $Q+1$ and that in Φ is Q, two strokes (for instance q^{th} and $q+1^{th}$) in Ψ should be coupled to one (q^{th}) to equalize both numbers.

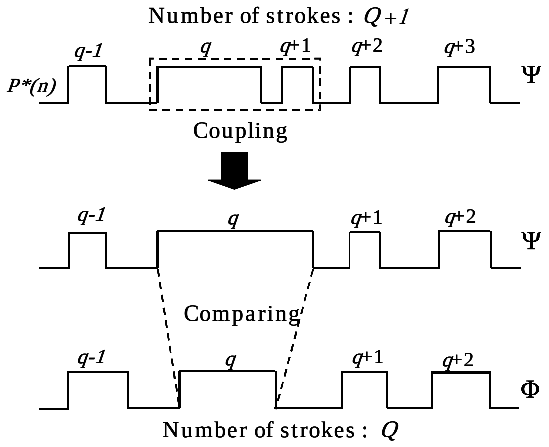


Figure 9. Stroke matching

It becomes important how to choose such strokes which should be coupled. Thus, we introduce the DP matching method, which is generally used for evaluating similarity between two patterns. In the following, we explain the DP matching method briefly.

The distance $d_q(i,j)$ between i^{th} sample of coupled q^{th} stroke in Ψ and j^{th} sample of q^{th} stroke in Φ is define as

$$d_q(i,j) = \left| t_{\Psi_q}(i) - t_{\Phi_q}(j) \right| \tag{16}$$

where $t_{\Psi_q}(i)$ and $t_{\Phi_q}(j)$ are normalized sampling times which are obtained by dividing whole writing time by total number of samples and different from those used in making of the template.

Next, the DP distance $D(\Psi_q,\Phi_q)$ between Ψ and Φ in q^{th} stroke is calculated as follows.

$$\textit{Initialize}: \quad g_q(0,0) = 2d_q(0,0) \tag{17}$$

For i=1 to I_q–1, j=1 to J_q–1

$$g_q(i,j) = \min \begin{bmatrix} g_q(i,j-1) + d_q(i,j) \\ g_q(i-1,j-1) + 2d_q(i,j) \\ g_q(i-1,j) + d_q(i,j) \end{bmatrix} \tag{18}$$

$$D(\Psi_q,\Phi_q) = \frac{g_q(I_q-1, J_q-1)}{I_q + J_q} \tag{19}$$

where I_q and J_q are the number of samples in q^{th} stroke of Ψ and Φ, respectively.

Such DP distance is calculated in all stroke pairs in Ψ and as a result a stroke pair with the shortest DP distance is chosen as the coupled stroke. In the case of $|\gamma|$=2, two couplings or a coupling of three strokes is examined similarly. Of course, in the case of γ=0, the stroke matching is not needed. If $|\gamma|$>2, the input signature is immediately decided as the forgery.

4.6 Adaptive signal processing for verification

After stroke matching, the verification is processed by using an adaptive signal processing. A block diagram of the proposed verification method by using the adaptive signal processing is shown in Fig.10. In this paper, we utilize *Detail*s at only *k*=*M*~*M*-*3*. *Detail*s at lower levels correspond to higher frequency elements and so their variation is too large. They are not suitable for verification. Input signals $x_k(n)$ and $y_k(n)$ are respectively the *Detail*s of x and y components at level *k* in an input signature. In the following, the signals of x and y components are represented as $v_k(n)$ (x, y∈v) for convenience. A desired signal $t_k^v(n)$ is the *Detail* of the template. $w_k^v(n)$ is an adaptive weight and updated to reduce the error signal $e_k^v(n)$ based on the adaptive algorithm (A.A.).

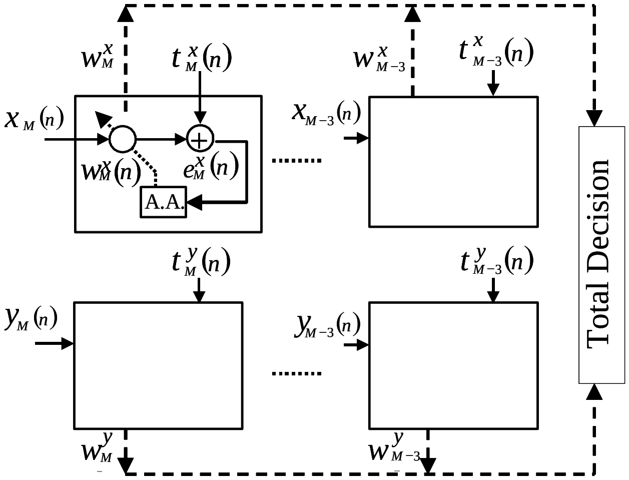


Figure 10. Adaptive signal processing for Verification

By the way, the purpose of the adaptive signal processing is to reduce the error between the input signal and the desired signal sample by sample. However, if these signals have different number of sampled data, the error does not fully decrease. In general, the number of data in the input signature does not necessarily agree with that in the template. In order to match these numbers of data, we utilize the normalized sampling time every intra/inter-stroke as described in 4.3. The nearest input data to the normalized sampled time in the template is only referred in the adaptive algorithm. Thus, the number of the input data is always guaranteed to agree with that in the template. Such time index according to the normalized sampled time is represented as *r*.

The adaptive algorithm for updating the weight is given by

$$w_k^v(r+1) = w_k^v(r) + \mu E\left[e_k^v(r) v_k(r)\right] \tag{20}$$

$$e_k^v(r) = t_k^v(r) - w_k^v(r) v_k(r) \tag{21}$$

$$E\left[e_k^v(r) v_k(r)\right] = \frac{1}{N}\sum_{l=0}^{N-1} e_k^v(r-l) v_k(r-l) \tag{22}$$

$$\mu = \frac{\mu_0}{\left\{E\left[\left| v_k(n) \right|\right]\right\}^2} \tag{23}$$

$$E\left[\left| v_k(n) \right|\right] = \frac{1}{L}\sum_{l=0}^{L-1} \left| v_k(n-l) \right| \tag{24}$$

where *L* is the number of sampled data in the input *Detail*. μ_0 is a positive constant and set to 0.0001, which is confirmed to be independent of the signature. μ is the step size parameter which controls the convergence in the adaptive algorithm. The step size parameter is normalized by the *Detail* power as shown in Eqs.(23) and (24), so that the convergence is always guaranteed. This algorithm is a kind of the steepest descent algorithm [7].

When the input signal is of the genuine signature, the error between the input and the template becomes small; therefore, the adaptive weight converges close on 1. Inversly, if the input signature is of the forgery, the weight converges far from 1. In this way, the verification can be achieved by examining whether the converged value is nealy 1 or not.

After enough iterations for convergence, the total convergence value *TC* is calculated by averaging eight converged values of the adaptive weights.

$$TC = \frac{1}{2}\left(\sum_{p=0}^{3}\frac{1}{4}\cdot w_{M-p}^{x} + \sum_{p=0}^{3}\frac{1}{4}\cdot w_{M-p}^{y}\right) \tag{25}$$

Finally, Total decision of verification is achieved by whether *TC* is larger than a threshold value. To consider multiple results leads robustness to our verification method.

5. Experiment and result

In order to confirm effectiveness of our proposed verification method, we carried out experiments of the signature verification. Conditions of the experiment are as follows. Four subjects were requested to sign their own alphabetic signature 25 times each and to counterfeit other two signatures 50 times each. Then, after excluding unusable signatures, 98 genuine signatures and 200 forgeries were used in this experiment. Before signing, subjects were called upon to practice using the pen tablet for becoming skilled. Also, when the subjects signed genuine signatures, they were not able to refer to their already written signatures. On the other hand, forgers were permitted to trace the genuine signature by putting the paper to which the signature was written over it. This assumed that the signature shape was easily imitated. In order to obtain fully convergence of the adaptive weight, the number of iterations was set to 100 thousands. In initial *N* iterations, the adaptive weights were fixed to zero and not updated since data for the average in Eq.(22) were uncompleted. In more than *N* iterations, the same input signal and template were used repeatedly.

Figure 11 shows an example of convergence characteristics of the adaptive weight of x component at level *M*. When the input signature was the genuine one, the adaptive weight converged nearer on 1 than that in the forgery. This result shows that it is possible to verify the signatures by using the adaptive signal processing.

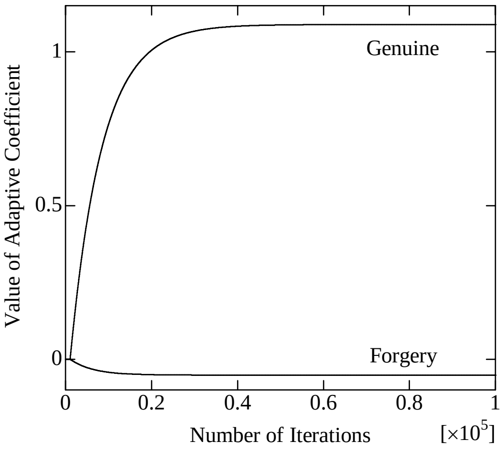


Figure 11. An example of convergence characteristics

Figure12 shows the false rejection rate (FRR) and the false acceptance rate (FAR) versus a threshold value. In general, verification performance is estimated by the equal error rate (EER) where the FRR and the FAR are the same. The EER was about 5% when the threshold value was set to about 0.3. This result means that the verification rate is 95% by using only the pen-position parameter even though a forger traces a genuine signature.

The reason why the FAR was not 100% even when the threshold value was set small enough was that about 20% forgeries had more than two stroke differences and so they were immediately rejected without verification.

On the other hand, the stroke difference of all genuine signatures was within ±2, so that the FRR became 100% as the threshold value was increased. The FRR did not become 0% even when the threshold was about 0 because the subjects were not permitted to refer to own signatures, and it enlarged variation in signature parameters. Some method for reducing the FRR is required to improve the verification rate.

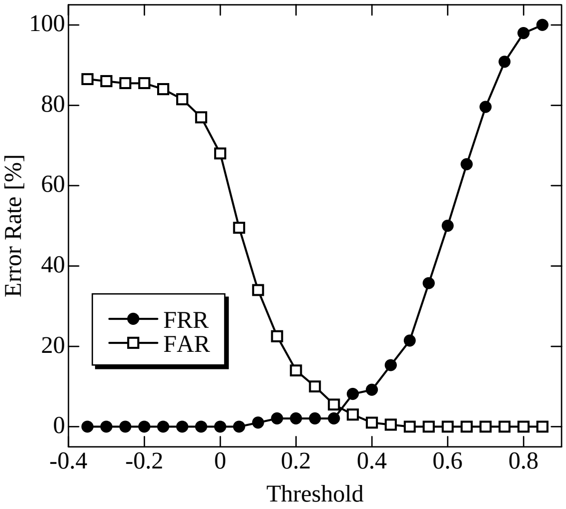


Figure 12. Verification results

6. Conclusion

We had proposed a new text-dependent on-line signature verification method. Our method emphasized individual features by decomposing the time-varying signal of the pen-position parameter into sub-band signals by the DWT. In addition, we proposed the verification method based on the adaptive signal processing, in which the normalized step size parameter was introduced to guarantee the convergence of the adaptive weight. Moreover, we adopted the DP matching method for stroke matching before the verification because stroke difference between signatures degraded verification performance.

Experimental results showed that the verification rate of 95% was achieved by using only the pen-position parameter under a very severe condition, that is, subjects were not able to refer to the genuine signatures and forgers were permitted to trace the genuine signatures. By using our proposed method, high verification rate can be achieved even in the portable device such as the PDA.

In the proposed method, the computational complexity is surely increased for the sub-band decomposition by the DWT, the adaptive signal processing for verification, and the DP matching method for stroke matching. In computational complexity, it is important to examine the number of multiplications and divisions. The DP matching is basically a repetition of addition and comparison as in Eq.(18); therefore, the increase of computational complexity is considered as not large. The adaptive signal processing is also based on iteration, and two multiplications and two divisions are required every iteration as in Eqs.(20), (21), (23) and (24). Eq.(22) denotes the moving average with window length *N*, and *N* multiplications, *N-1* additions and one division are needed every iteration. However, *N-1* data in accumulation are overlapped between a present window and a past window since the window is shifted sample by sample. By adding a present data to the accumulation and subtracting a past data from the accumulation, *N* multiplications can be reduced to one. As a result, three multiplications and three divisions are required every iteration. The increase of the computational complexity is not also large.

In this paper, converged values of adaptive weights are simply averaged to obtain the total convergence value. To adjust weighting of the converged value should be introduced for improving verification performance. For reducing the FRR, it is to also studied in future to cope with variation in the genuine signature.

7. References

[1] A. Jain, R. Bolle and S. Pankanti, *BIOMETRICS Personal Identification in Networked Society*, Kluwer Academic Publishers, Massachusetts, 1999.

[2] Y. Sato and K. Kogure, "Online Signature Verification Based on Shape, Motion, and Writing Pressure," *Proc. of 6th Int. Conf. on Pat. Recog.*, 1982, pp.823-826.

[3] M. Yoshimura, Y. Kato, S. Matsuda, and I. Yoshimura, "On-line Signature Verification Incorporating the Direction of Pen Movement," *IEICE Trans.*, vol.E74, no.7, July 1991, pp.2083-2092.

[4] Y. Yamazaki and N. Komatsu, "Extraction of Personal Features from On-Line Handwriting Information in Context-Independent Characters," *IEICE Trans. Fundamentals*, vol.E83-A, no.10, Oct. 2000, pp.1955-1962.

[5] Y. Komiya, T. Ohishi and T. Matsumoto, "A Pen Input On-Line Signature Verifier Integrating Position, Pressure and Inclination Trajectories," *IEICE Trans. Inf. & Syst.*, vol.E84-D, no.7, July 2001, pp.833-838.

[6] G. Strang, T. Nguyen, *Wavelet and Filter Banks*, Wellesley-Cambridge Press, 1997.

[7] S. Haykin, *Introduction to Adaptive Filters*, Macmillan publishing Company, New York, 1984

Multimodal Speaker Authentication using Nonacoustic Sensors[*]

W. M. Campbell, T. F. Quatieri, J. P. Campbell, C. J. Weinstein
MIT Lincoln Laboratory
{wcampbell, tfq, jpc, cjw}@ll.mit.edu

Abstract

Many nonacoustic sensors are now available to augment user authentication. Devices such as the GEMS (glottal electromagnetic micro-power sensor), the EGG (electroglottograph), and the P-mic (physiological mic) all have distinct methods of measuring physical processes associated with speech production. A potential exciting aspect of the application of these sensors is that they are less influenced by acoustic noise than a microphone. A drawback of having many sensors available is the need to develop features and classification technologies appropriate to each sensor. We therefore learn feature extraction based on data. State of the art classification with Gaussian Mixture Models and Support Vector Machines is then applied for multimodal authentication. We apply our techniques to two databases—the Lawrence Livermore GEMS corpus and the DARPA Advanced Speech Encoding Pilot corpus. We show the potential of nonacoustic sensors to increase authentication accuracy in realistic situations.

1. Introduction

Speaker authentication is a rich area for exploration of multimodality. Many facets of the speech production process are measurable through a variety of sensors. Traditionally, visual lip reading has been used to supplement speaker authentication and speech recognition [15,26]. These methods rely upon tracking the lip contour over time and then using the sequence of movements to supplement standard audio-only verification. These methods have been quite successful, leading to large gains in accuracy in high noise conditions.

Other methods of monitoring speech production are also available. Non-invasive sensors that are attached in the throat area have been available for many years; we call these nonacoustic sensors. These sensors nominally measure aspects of the speech production process related to the speech excitation. Typical sensors that we have explored in this study are the EGG (electroglottograph), the GEMS (glottal electromagnetic micro-power sensor), and the P-mic (physiological mic). Since traditional methods of verification [18] rely upon features designed to capture vocal tract information—e.g., mel-frequency cepstral coefficients—we would expect that multimodal fusing of excitation and vocal tract features would benefit recognition in *both* quiet and noisy conditions. An added benefit of nonacoustic sensors is that they are less influenced by acoustic noise. For the case of the EGG and the GEMS, the throat is exposed to RF signals; for the case of the P-mic, the sensor output is dominated by the vibrations sensed on the throat. These modes of measurement do not directly monitor air pressure in the ambient environment.

There has been several prior works on the use of glottal waveforms for recognition. Gable [8] used waveforms from the GEMS system for speaker verification; his work focused on using methods such as dynamic time warping for text-dependent verification. Plumpe [16] used inverse filtering techniques on the acoustic waveform to derive glottal waveform signals; speaker recognition was then performed. Both throat microphones [9] and the P-mic [1] have been used for automatic *speech* recognition. Our work is distinct in several aspects: 1) we consider both simulated and actual noise conditions, 2) we do not assume models for the glottal waveforms but instead use a learning approach, 3) we use late integration to combine *several* nonacoustic sensors, and 4) we consider integration accuracy of multiple nonacoustic sensors in low-noise conditions.

We attack the problem of authentication using nonacoustic sensors with a data-driven learning approach. We have chosen the data-driven approach as a baseline to future knowledge-based analysis. Sensor outputs can vary dramatically based on placement, sensor tuning, impedance matching, sensor design, etc. This variation can be captured easily with data-driven methods. Towards this end, we use standard feature transformation methods to find features which describe the speaker specific attributes of the different signals. We use various normali-

[*] This work is sponsored by the Defense Advanced Research Projects Agency under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

zations based upon signal characteristics to improve accuracy.

After obtaining features for authentication, we use both Gaussian Mixture Models [18] and Support Vector Machines (SVM's) [25] for multimodal authentication. We combine the outputs of these different classification systems using late integration to achieve the final score. For the corpora explored in this paper, we consider only closed-set speaker identification. That is, given an utterance, identify an individual from a list of known individuals. Because of the limited number of speakers available in current corpora, other scenarios such as verification or open-set ID were impossible because of the lack of an adequate "background" population.

The outline of the paper is as follows. In Section 2, we discuss the sensors in detail and describe their basic operation. In Section 3, we discuss our feature extraction methodology. Section 4 outlines the classifiers and fusion strategy used. Section 5 gives details on the corpora used and experiments. These corpora allow us to explore both the GEMS in quiet environments and multiple nonacoustic sensors in high noise (>110 dBC) situations. We show that our authentication strategy leads to gains in this challenging scenario. A complimentary method for achieving authentication accuracy gains is speech enhancement [27].

2. Nonacoustic sensors

We survey three nonacoustic sensors used for experiments—GEMS, EGG, and P-mic. These sensors have distinct methods of measuring speech production phenomena. Other sensors which would be of interest, but were not included due to corpus size and project focus, are accelerometers, "bone phones," in-ear microphones, video, etc.

2.1. GEMS

The GEMS (glottal electromagnetic micro-power sensor) is a novel sensor based upon transmitting electromagnetic (EM) waves into the glottal region. Two GEMS designs were used in the corpora in this paper. An earlier version was used in the LLNL Corpus [8], and Revision B, Version 1 created by Aliph Corporation (http://www.aliph.com) was used in the ASE Corpus of Section 5. The GEMS is also referred to as the "General Electromagnetic Movement Sensor" by Aliph Corporation.

During operation of the GEMS, a small antenna is placed on or near the throat at the level of the glottis. From this antenna is transmitted a 2.3 or 2.4 GHz low power (<1 mW) EM wave. Using these frequencies allows for EM waves to penetrate into the body and reflect back to the sensor with good signal levels. The receiver circuitry detects the reflected EM waves using a homodyne technique. Nominally, the sensor measures phenomena related to the opening and closing of the glottis [2]. Multiple theories have emerged on the exact phenomena occurring that generates the waveform—changing air-tissue interfaces as the glottis changes, vibration of the tracheal wall, and propagation along the vocal fold contact area, see [11, 21]. Although inferring the exact process that the GEMS is monitoring is challenging, the waveforms generated do provide speaker specific information which is related to the speech excitation.

2.2. EGG

The EGG (electroglottograph) is a device designed to measure contact between the vocal folds. The specific implementation used for this study was from Glottal Enterprises. This EGG is a multi-channel EGG device [19]; the multichannel feature allows for more precise placement on the neck to achieve higher signal to noise ratio.

The EGG nominally measures the vocal fold contact *area* (VFCA). This process is performed by using electrical signals in the MHz region. Two electrodes are placed on the subject's neck at the level of the thyroid cartilage. VFCA is measured by observing the variation in impedance over time. Since the EGG measures vocal fold contact, the sensor does not necessarily allow one to observe interesting phenomena during the open phase of the glottis. Note that the EGG is not an exact indicator of VFCA. For example, during transition to the open phase of the glottis, mucus can "short out" the device indicating that the glottis is closed when this is apparently not the case (the mucus bridging effect [4]).

2.3. P-mic

The P-mic (physiological microphone) is a non-invasive contact sensor for measuring sound [20]. The P-mic consists of a gel pad to provide acoustic impedance matching, a conical focusing aperture, and a piezoelectric element. Use of a gel pad minimizes interference from ambient noise.

The P-mic is typically placed in the throat area below the glottis. This placement insures that the P-mic signal can be simultaneously recorded with the GEMS and EGG signal. In our experiments, we found that the P-mic was most sensitive to ambient noise among nonacoustic sensors; presumably this is due to "leakage" of the ambient noise into the sensor element.

2.4. Comparison of the sensors

Figure 1 shows an example output from four sensors recorded simultaneously. In the figure, the top signal is a microphone recording of the /ao/ in "dog." The second signal represents the EGG signal (highpass filtered with a linear phase filter with a transition band from 64-80 Hz). We note that the EGG gives a very "smooth" waveform. The third waveform from the top is the P-mic signal. In this signal, we see more evidence of "leakage" of vocal tract information into the signal (as evidenced by ripple in the waveform). Finally, the fourth waveform is the GEMS signal. We can see this waveform has many of the same general characteristics as the EGG, but that there is additional structure in the waveform. Listening to the GEMS signal reveals little vocal tract information; therefore, this fine structured seems to represent supplementary excitation information not captured by the EGG.

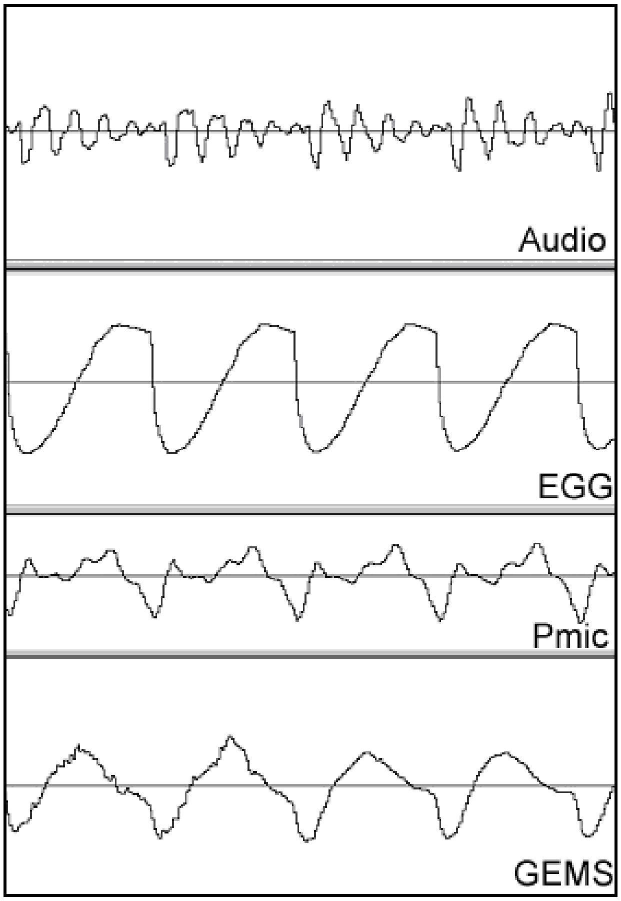


Figure 1. Comparison of different sensor waveforms for the /ao/ in "dog." From top to bottom—audio, EGG, P-mic, and GEMS. The length of time shown is approximately 30 ms.

3. Feature extraction

Our framework for feature extraction is shown in Figure 2. Our goal was to create a flexible architecture that incorporated linear matrix transformation for feature extraction. In the figure, the input signal is processed into frames creating a sequence of vectors. Each frame corresponds to a 30 ms time window with an overlap of 20 ms between consecutive frames. Since our sampling rate is 8 kHz, we obtain a sequence of vectors of dimension 240 (100 vectors per second).

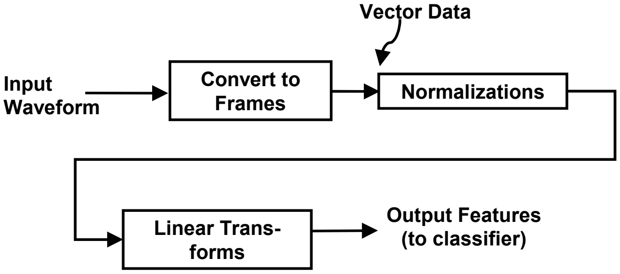


Figure 2. Framework for feature transformation.

We then applied several normalizations to the data; these normalizations are intended to provide invariances in the feature extraction to certain transforms—e.g., increasing the gain. We first remove the mean on a per frame basis; we then normalize the amplitude of the signal variance to 1. Finally, we introduce a transform to reduce a framing artifact; namely, a shift of the input should not matter in recognition. For this normalization, we calculate the discrete Fourier transform (DFT) of each frame, eliminate the phase of each component, and then calculate the inverse DFT. All of these normalizations are intended to throw out unnecessary signal information; potentially, they are too aggressive and could be modified. For example, the mean of the EGG signal carries information about the position of the larynx. In spite of drawbacks, these normalizations increased accuracy for all linear transforms we tried.

After appropriate normalization, the sequence of frames was used to calculate delta parameters [17]. This linear transform resulted in a sequence of vectors of dimension 480. We then wanted to design a linear transform to reduce this 480 component vector to a more reasonable dimension. There are multiple reasons for dimension reduction—obtaining compact representations of speaker specific features, avoiding excessively complex classifiers, discarding "uninformative" directions in feature space, and minimizing the "curse of dimensionality." For this paper, we explored several unsupervised methods of designing a linear transform—principal component analysis (PCA) [7], random dimension reduction [6], and independent component analysis (ICA) [12].

Random dimension reduction (i.e., generating the analysis matrix using random independent components) was used for multiple purposes. We preprocessed all of the normalized outputs (with delta components) from dimension 480 down to dimension 100 using random dimension reduction. As shown in [6], random dimension reduction tends to preserve distances and make clusters of data more spherical which improves problem conditioning. We found that for both PCA and ICA that this improved accuracy. Random dimension reduction also reduces the size of the problem making methods such as ICA and PCA more practical for large problems. Finally, random dimension reduction was also used as an analysis method to compare to other unsupervised methods.

We note that our feature transformation method is very similar to the standard filter bank approach for generating mel-cepstral coefficients. In a coarse sense, our approach could be thought of as applying a filter bank "tuned" to the glottal response.

4. Classification and fusion

Gaussian mixture models have been very successful for the speaker recognition task [18]. We use Gaussian mixture models to model the speaker specific distribution only (i.e., no background modeling is performed since our task is closed-set identification). For each speaker, we create a mixture model

$$f(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i g_i(\mathbf{x})$$

where g_i is a single Gaussian. Training is accomplished using the EM algorithm with a small number of components—typically less that 256.

We also use support vector machines (SVM's) for classification [25]. Support vector machines are discriminatively trained classifiers and thus give excellent performance on closed set tasks. For our experiments, we use a polynomial basis of monomials in our SVM kernel up to and including a certain degree—typically degree 2 or 3, see [25]. Our SVM kernel is based upon comparing sequences of data and providing an inner product in a large dimensional space which captures speaker specific information. One interesting aspect of using support vector machines for our work is that it is possible to bypass the feature transformation process and perform classification directly in high dimensions. Although this is computationally intense, it gives a baseline for feature transformed classification systems which work in lower dimensions.

All of our reported experiments use late integration for fusion [3]. Fusion is accomplished by using a linear combination of scores from each of the classifiers applied to the different modalities. Methods involving construction of new SVM kernels based upon sums of kernels for each of the modalities were also tried, but these did not perform as well as late integration.

5. Corpora and experiments

5.1. LLNL GEMS corpus and experimental setup

The first corpus used for experiments was the Lawrence Livermore National Lab GEMS corpus collected by G. Burnett and T. Gable [8]. This corpus consists of 15 male speakers with up to 4 sessions per speaker. Both sentences from TIMIT and number/letter/{Yes,No,Zero} sequences were recorded. For the purposes of our experiments, we focused on the number/letter/short-word sequences. Typical utterances were a combination of 10 items; e.g., "T 60 YES 3 U R E 8 W P."

We used the initial session of 20 utterances as enrollment. The remaining 3 sessions of 20 utterances each were used for speaker identification. This resulted in 15*60=900 tests for speaker identification. Both audio and GEMS data were originally sampled at 10 kHz. We resampled to 8 kHz and then bandlimited the speech to 200-4000 Hz.

Noise was electronically added to the audio signal with noises from the NOISEX database [23]. (In Section 5.3 and 5.4, we consider a corpus where the noise environment is not electrically added.) The NOISEX noise signals were resampled to 8 kHz and also bandlimited to 200-4000 Hz. This insured that SNR was measured only in the band containing speech. All 24 NOISEX noises were used. When adding speech to noise, we generated a random offset into the noise file and then extracted a segment of noise the same length as the speech file. The resulting output signal was $x = x_{\text{speech}} + c * x_{\text{noise}}$, where

$$c = \frac{\sigma_{speech}}{\sigma_{noise}} 10^{-\frac{\text{SNR}}{10}}$$

and the standard deviations are calculated over non-silence regions.

5.2. LLNL corpus results

Our first set of experiments compared feature transformation methods. As indicated in Section 4, we explored random dimension reduction, PCA, and ICA. We initially considered closed-set speaker identification accuracy based upon the GEMS signal only. Each feature vector was reduced from dimension 480 to 100 using

Table 1. Comparison of accuracy of feature transformation methods for GEMS-only closed-set speaker identification on the LLNL database.

Feature Extraction Method	Speaker Identification Accuracy (%)
Random Projection	62.7 %
PCA	59.7 %
ICA	51.9 %
None	64.3 %

random dimension reduction. A linear transform was then designed and applied to reduce the dimension from 100 to 32 for input to the classifier. Dimension 32 was chosen since the accuracy typically plateaued at this dimension. A SVM classifier with a degree 2 polynomial kernel (full covariance) was used, see [25].

Table 1 compares accuracies for the different methods. Also included in the table is the case of no dimension reduction (with a diagonal covariance SVM kernel) which provides a baseline for reduced dimension methods. As can be seen from the table, random projection works as well as other transformation methods. Potentially, this is due to multiple factors. The classifier may be better matched to this feature extraction technique. Also, there could be spurious directions in the feature space data which are not relevant to speaker identification. One way to mitigate this problem (which we do not explore here) is to use supervised feature transformation methods, e.g. [22].

After using linear transform feature extraction methods for speaker identification, we investigated the use of fundamental frequency (F0) to augment the recognition process. The Entropic pitch extractor in Wavesurfer (http://www.speech.kth.se/wavesurfer) was used. A GMM was trained with 32 components to model each speaker from the F0 data. The resulting error rate for GEMS only recognition was 50.6%. Note that a similar rate of accuracy was also observed for the audio data using F0 only—49.1%.

We then fused (with equal weights) the GEMS F0 classifier scores with the linear transform feature extraction scores (random dimension reduction) to obtain a GEMS-only accuracy of 64.0%. The use of F0 information demonstrated two items. First, since F0-only classification accuracy is significantly below that of linear transform feature extraction accuracy, we are obtaining additional non-F0 information from our linear transform technique.

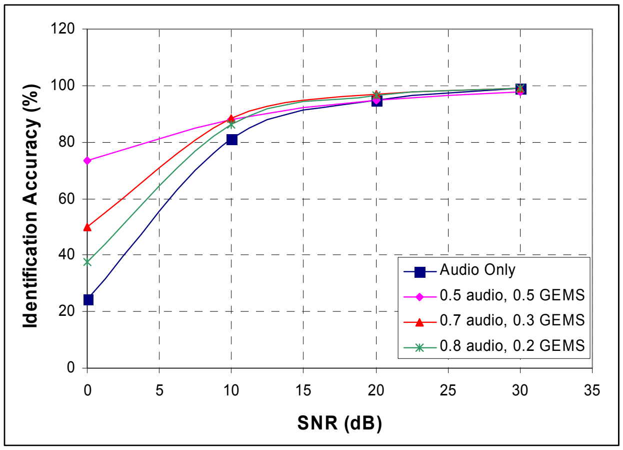


Figure 3. Comparison of speaker identification accuracy across noise type 3 (white noise) for different late integration strategies and random dimension reduction.

Second, because the accuracy improved from the fusion, there is complementary information in the two scores.

Finally, we considered the effect of late integration upon speaker identification in noise. We implemented an audio-only speaker recognition system using the system in [25] with a degree 3, diagonal covariance model; input features were 12 LP cepstral coefficients plus deltas. In addition, the MELPe noise preprocessor [24] was applied to the audio input signal. Figure 3 shows the performance of a late integration system which fuses an audio-based system with the GEMS-based system (both pitch and linear feature transformation were used). In the figure, at low SNR (0-10 dB) and for NOISEX white noise (noise type 3), significant increases in accuracy are obtained by late integration—greater than 50% in some cases.

We then considered the effect of late integration with a fixed weighting, 0.5*GEMS + 0.5*audio, as the type of noise varied for a fixed SNR (specific information on the noise types can be found in the NOISEX corpus documentation). The results for 0 dB SNR are shown in Figure 4. As can be seen from the figure, significant increases in accuracy over an audio-only system are achieved—greater than 25% average improvement. The best performing environments were NOISEX types 3 (white noise), 16 (machine gun), 18 (STI test signal), 19 (voice babble), and 21 (factory). The worst performing environments were NOISEX types 1 (sinusoid), 5 (colored, -12 dB/octave), 9 (Leopard 2), 23 (Car) and 24 (Car).

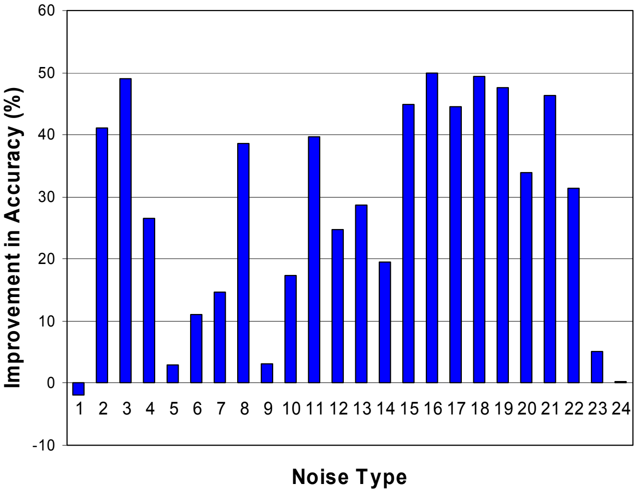


Figure 4. Improvement in speaker identification accuracy of a late-integration fusion system over an audio-only system by noise type (NOISEX database) at 0 dB SNR.

5.3. ASE corpus and experimental setup

The Advanced Speech Encoding Pilot Corpus (ASE Pilot Corpus) is a multisensor corpus collected for the purpose of studying viability of multiple sensors for speech enhancement, speech coding, and speaker characterization. Sensors recorded simultaneously include a resident microphone (the microphone typically used in the environment), two channels of a GEMS device, an EGG, a high quality reference microphone (B&K), and P-mics positioned on the forehead and the throat region. The corpus was collected in two sessions (on two different days). Speakers were exposed to a variety of noise environments—-quiet, office (56 dBC), MCE (mobile command enclosure, 79 dBC), M2 Bradley Fighting Vehicle (74 dBC and 114 dBC), MOUT (military operations in urban terrain, 73 dBC and 113 dBC), and a Blackhawk helicopter (70 dBC and 110 dBC). We call these environments (with L indicating low noise and H indicating high noise) quiet, office, MCE, M2L, M2H, MOUTL, MOUTH, BHL and BHH, respectively. To protect our subjects and realistically simulate Lombard effects, all talkers used the hearing protection systems typical of each environment. This normally consisted of a communication headset with approximately 20 dB noise attenuation. Human subject testing procedures were followed carefully and noise exposure was monitored.

For speaker identification experiments, we partitioned the corpus by session. The initial sessions—quiet, office, and MCE—were used for enrollment. Identification was then performed using the data from the remaining sessions; we grouped these into low noise—M2L, MOUTL, BHL—and high noise—M2H, MOUTH, BHH—conditions. The corpus had phrases in both sessions drawn from a variety of material—conversations, DRT lists, vowels, Harvard phonetically balanced sentences, and CVC nonsense words. Typical utterance lengths ranged from 1-5 minutes. A total of 20 speakers were available, 10 males and 10 females. The total number of enrollment utterance available per speaker was 12. The total number of tests for identification performance was 360 per noise condition (low, high). Cross-gender testing was allowed since it was not clear if the nonacoustic sensors would distinguish this well; cross-gender tests do not bias identification accuracy (as they would in speaker verification).

5.4. ASE corpus results

The feature extraction methods from Section 3 were applied to the ASE pilot corpus. As for the experiments in Section 5.2, we used a SVM with diagonal covariance and degree 3 polynomials for the audio modality. For the nonacoustic modalities, we used a full covariance SVM of degree 2 with random dimension reduction. Both the MELPe noise preprocessor and high-pass filtering above 200 Hz were applied to the audio signal. The MELPe noise preprocessor was applied to the non-acoustic modalities, since noise from the ambient environment did effect the sensor outputs (possibly through tissue vibration). The EGG was highpass filtered with a linear phase filter with transition band from 64-80 Hz. Results are shown in Table 2.

Since the P-mic has some vocal tract information (as evidenced by listening), we also applied a standard LP cepstral coefficient front end to the data; i.e., we applied the audio recognition system to all sensors. Results for this set of experiments are shown in Table 3. As can be seen from the table, accuracy results for both the EGG and GEMS are generally lower for LP cepstral coefficients than with data driven methods shown in Table 2. For the P-mic, the identification accuracy is higher for LPCC's; this illustrates that standard methods are tuned to extracting vocal tract information.

Table 2. Identification accuracy in both low and high noise situations for multiple modalities using random dimension reduction.

Modality	Low Noise Accuracy	High Noise Accuracy
EGG	73.0 %	43.3 %
GEMS	64.7 %	43.6 %
P-mic	66.7 %	41.4 %

Table 3. Identification accuracy in both low and high noise situations for multiple modalities using LP cepstral coefficients.

Modality	Low Noise Accuracy	High Noise Accuracy
Resident Mic	89.4 %	81.9 %
EGG	61.1 %	38.0 %
GEMS	50.3 %	43.6 %
P-mic	77.5 %	55.0 %

Table 4. Identification accuracy in both low and high noise situations for late integration fusion.

Modalities Fused	Low Noise Accuracy	High Noise Accuracy
Audio (Resident Mic)	89.4 %	81.9 %
0.8*Audio+0.2*EGG	93.1 %	86.7 %
0.8*Audio+0.2*GEMS	92.5 %	85.8 %
0.5*Audio+0.5*P-mic	95.8 %	87.2 %
All	95.8 %	89.4 %

Two items should be noted about the results in Tables 2 and 3. First, the accuracy of the resident microphone is somewhat low in low noise situations. This result is probably due to mismatch in microphones between training and testing. Second, high-noise accuracy of the resident microphone is quite good. The MELPe noise preprocessor and associated processing is fairly robust to noise.

Another observation from Tables 2 and 3 is the degradation of nonacoustic sensors in noise. For the GEMS modeling in Section 5.2, we assumed the ideal case of no degradation due to noise. It is well known in the literature [5], that even if acoustic noise is not present in the sensor data, a human speaker responds to the environment, e.g. Lombard effect [13]. This response to stress will cause degradation in the speaker identification performance of the nonacoustic modalities. An open research question is how to compensate for the effects of stress in the excitation parameterization. Although we do not explore methods here, the ASE pilot corpus provides a realistic scenario for studying methods of noise compensation of the speech excitation waveform.

Table 4 shows the results of late integration. For the EGG and GEMS, fusion with the weights shown and random dimension reduction yielded the best results. For the P-mic, LPCC's performed the best with equal weighting of audio and P-mic modalities. For the fusion of all modalities, we tried a variety of weightings; the best performing weighting was 0.5*audio, 0.2*EGG, 0*GEMS, and 0.3*P-mic (labeled "All" in Table 4). Unfortunately, a cross-validation data set was not available to validate the fusion process.

As indicated in Table 4, we obtain substantial gains of 7.5% in speaker identification accuracy in noise, over the resident-microphone-only case by combining nonacoustic and acoustic scores. This result shows the potential of these methods for noise robust speaker authentication.

6. Conclusions

We have demonstrated the use of nonacoustic sensors for speaker authentication. A data-driven approach was used to derive features of different modalities. Powerful classification techniques such as support vector machines and Gaussian mixture models were then applied. Results in both simulated and actual noisy conditions showed the success of the techniques for dramatically improving speaker authentication in noise. Future work should explore methods on statistically-significant larger speaker populations to further validate results.

Acknowledgements

We thank John Tardelli and Paul Gatewood of ARCON Corporation for their excellent work in collecting the ASE multisensor corpus used for experiments in this paper. We thank Kevin Brady of MIT Lincoln Laboratory for his extensive support in this corpus collection. We thank Doug Reynolds for consultation on speaker recognition.

References

[1] Bass, J. D., M. V. Scanlon, T. K. Mills and J. J. Morgan, "Getting two birds with one phone: an acoustic sensor for both speech recognition and medical monitoring," *presentation at 138th meeting of the Acoustical Society of America, Columbus, OH,* 1999.

[2] Burnett, G. C., The physiological basis of Glottal Electromagnetic Sensors (GEMS) and their use in defining an excitation function for the human vocal tract, PhD Thesis, University of California, Davis, 1999.

[3] Chen, T. and R. R. Rao, "Audio-Visual integration in multimodal communication," *Proceedings of the IEEE*, 1998, pp. 837-852.

[4] Childers, D.G and A. K. Krishnamurthy, "A critical review of electroglottography", CRC Critical Reviews in Biomedical Engineering, 12, 1985, pp. 131-161.

[5] Cummings, K. E. and M. A. Clements, "Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering," *Proceedings Southeastcon*, 1989, pp. 776-781.

[6] Dasgupta, S., “Experiments with Random Projection,” *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 2000, pp. 143-151.

[7] Fukunaga, K., *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., San Diego, CA, 1990.

[8] Gable, T.J., *Speaker Verification Using Acoustic and Glottal Electromagnetic Micro-power Sensor (GEMS) Data*, PhD Thesis, University of California, Davis, 2000.

[9] Graciarena, M., H. Franco, K. Sonmez and H. Bratt, “Combining standard and throat microphones for robust speech recognition,” *IEEE Signal Processing Letters,* vol. 10, no. 3, 2001, pp. 72-74.

[10] Holzrichter, J. F., G. C. Burnett, L. C. Ng and W. A. Lea, “Speech articulator measurements using low power EM-wave sensors,” *Journal of the Acoustical Society of America*, 1998, 103(1), pp. 622-625.

[11] Holzrichter, J. F., L. C. Ng, G. J. Burke, N. J. Champagne II, J. S. Kallman, R. M. Sharpe, J. B. Kobler, R. E. Hillman and J. J. Rosowski, “EM wave measurements of glottal structure dynamics,” *University of California, Lawrence Livermore Laboratory Report*, UCRL-JC-147775 , 2002.

[12] Hyvarinen, A., “Fast and Robust Fixed-Point Algorithms for Independent Component Analysis,” *IEEE Trans. On Neural Networks*, vol. 10, no. 3, 1999, pp. 626-634.

[13] Lippmann, R. P., E. A. Martin, “Multi-Style Training for Robust Isolated-Word Speech Recognition,” *Proceedings IEEE International Conference on Acoustics Speech and Signal Processing*, 1987, pp. 705-708.

[14] Lawrence Livermore National Lab, Glottal Electromagnetic Micropower Sensor and Acoustic Data, http://speech.llnl.gov, 1999.

[15] Luettin, J., N. Thacker and S. Beer, “Speaker Identification by Lipreading,” *Proc. ICSLP*, 1996, pp. 62-64.

[16] Plumpe, M. D., T. F. Quatieri and D. A. Reynolds, “Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identfication,” *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 5, 1999, pp. 569-586.

[17] Rabiner, L. and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[18] Reynolds, D. A., “Speaker Identification and Verification using Gaussian Mixture Speaker Models,” *Speech Communication*, vol. 17, 1995, pp. 91-108.

[19] Rothenberg, M. “A Multichannel Electroglottograph,” *J. of Voice*, 1992, vol. 6, no.1, pp. 36-43.

[20] Scanlon, M. V., “Acoustic Sensor for Health Status Monitoring,” *Proceeding of IRIS Acoustic and Seismic Sensing*, 1998, Volume II, pages 205-222.

[21] Titze, I. R., B. H. Story, G. Burnett, J. F. Holzrichter, L. C. Ng, W. A. Lea, “Comparison between electroglottography and electromagnetic glottography,” *J. Acoust. Soc. Am.*, vol. 107, no. 1, 2000, pp. 581-588.

[22] Torkkola, K. and W. Campbell, “Mutual information in learning feature transformations,” *Seventeenth International Conference on Machine Learning*, 2000, pp. 1015-1022.

[23] A.P. Varga, H.J.M Steenekan, M. Tomlinson, and D. Jones, “The NOISEX-92 study on the effect of additive noise on automatic speech recognition,” *Tech. Rep., DRA Speech Research Unit*, 1992.

[24] Wang, T., K. Koishida, V. Cuperman, A. Gersho and J. S. Collura, "A 1200/2400 BPS Coding Suite Based on MELP," NATO AC/322(SC/6-AHWG/3) AD HOC Working Group on Narrow Band Voice Coding, 2002 IEEE Workshop on Speech Coding, Special Session 1: Topics on NATO Standardization, Tsukuba, Ibaraki, Japan, October 6-9, 2002.

[25] Campbell, W. M., “Generalized Linear Discriminant Sequence Kernels for Speaker Recognition,” *Proceedings of ICASSP*, 2002, pp. 161-164.

[26] Zhang, X., C. C. Broun, R. M. Mersereau and M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces," *Eurasip Journal on Applied Signal Processing*, 2002, pp. 1228-1247.

[27] Quatieri, T. F., D. Messing, K. Brady, W. M. Campbell, J. Campbell, M. Brandstein, C. Weinstein, J. Tardelli, and P. Gatewood, “Exploiting non-acoustic sensors for speech enhancement,” *submitted to Workshop on Multimodal User Authentication.*

Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition

Douglas Reynolds, Joe Campbell, Bill Campbell, Bob Dunn, Terry Gleason, Doug Jones,
Tom Quatieri, Carl Quillen, Doug Sturim, Pedro Torres-Carrasquillo

MIT Lincoln Laboratory
244 Wood St
Lexington, MA 02420
{dar,jpc,wcampbell,rbd,tpg,daj,tfq,cbq,sturim,ptorres}@ll.mit.edu

Abstract

Traditionally speaker recognition techniques have focused on using short-term, low-level acoustic information such as cepstra features extracted over 20-30 ms windows of speech. But speech is a complex behavior conveying more information about the speaker than merely the sounds that are characteristic of his vocal apparatus. This higher-level information includes speaker-specific prosodics, pronunciations, word usage and conversational style. In this paper, we review some of the techniques to extract and apply these sources of high-level information with results from the NIST 2003 Extended Data Task.

1. Introduction

Standard approaches to automatic speaker recognition have relied on using short-term acoustic features, such as cepstra, which convey information about the shape of a person's vocal apparatus. While these approaches have shown success, speech is the product of a complex behavior conveying many other person-specific traits that are potential sources of complementary information. Roughly we can categorize information in speech into a hierarchy running from low-level information, such as the sound of a person's voice, which is related to physical traits of the vocal apparatus, to high-level information, such as particular word usage (idiolect), conversational patterns and even topics of conversations, which is related to learned habits and style (see Figure 1).

With the continual improvement of phoneme and speech recognition systems, which can reliably extract features for high-level characterization, the widespread availability of the computational resources needed to train and run them, and finally with the increased focus on applications (like audio mining) allowing for relatively large amounts of speech from a speaker to learn speaking habits, the availability of large development corpora and plentiful computational resources, the time is right for a deeper exploration into using these underutilized high-level information sources. These new sources of information hold the promise not only for improvement in basic recognition accuracy by adding complementary knowledge, but also the possibility for robustness to acoustic degradations from channel and noise effects, to which low-level features are highly susceptible. Over the last few years, work examining the exploitation of high-level information sources, such as the SuperSID Project [1][i], has provided strong evidence that gains are possible.

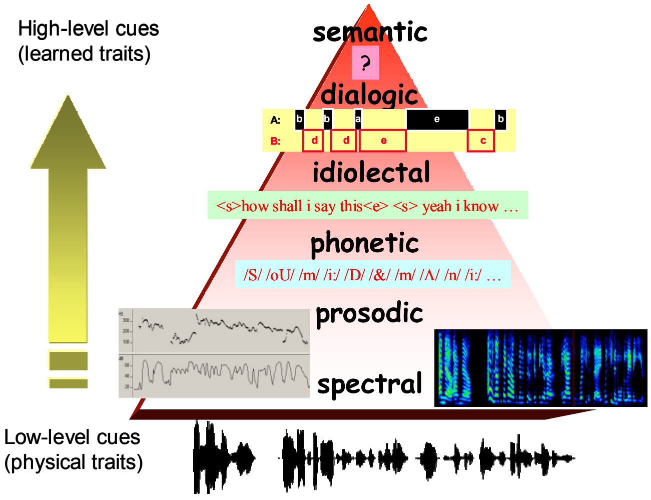


Figure 1 Pictorial depiction of levels of information conveyed in speech

To support exploration and development of techniques that can exploit large amounts of training data to learn speaker habits, an Extended Data Task (EDT) has been

This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government. After first author, subsequent authors are listed alphabetically.

[i] More references available at the SuperSID Project website http://www.clsp.jhu.edu/ws2002/groups/supersid/

included in the annual NIST Speaker Recognition evaluations since 2001. In this paper, we review some of the techniques to extract and apply these sources of high-level information with results from the NIST 2003 Extended Data Task. In the next section, we describe the corpus of the extended data task. This is followed in Section 3 by an outline of the different features and systems applied. Section 4 presents results of the systems on the 2003 EDT showing how they can be fused to improve overall performance.

2. The 2003 NIST Extended Data Task

The focus of the NIST SRE is on text-independent speaker detection (verification) using telephone speech. The primary evaluation has generally been using two minutes of training data and 15-45 seconds of test data. In 2001 the extended data task was introduced to allow exploration and development of techniques that can exploit significantly more training and testing data. Speaker models are trained using 1, 2, 4, 8, and 16 complete conversation sides (where a conversation side nominally contains 2.5 minutes of speech). A complete conversation side was used for testing. The 2003 Extended Data Task used the combined Switchboard-II phase-2 and phase-3 conversational telephone speech corpora.

To supply a large number of target and nontarget trials and speaker models trained with up to 16 conversations of training speech (~40 minutes), the evaluation used a jackknife processing of the entire corpus. The corpus was divided into 10 partitions of ~106 speakers each. All trials within a partition involved models and test segments from within that partition, only; data from the other 9 partitions were available for background model building, normalization, etc. The task consists of 1065 speakers with 10,933 target models (a speaker had multiple models for different amounts of training data) and ~160,000 trials (36% target trials and 64% nontarget trials) for the testing phase, containing matched and mismatched handset trials and some cross-sex trials. The experiments were driven by NIST's speaker model training lists and index files indicating which models were to be scored against which conversation sides for each partition.

To help facilitate research into using new features, supplemental information contributed by various sites was made available by NIST. This includes automatically generated word level transcripts, phone level transcripts from five different language phone sets, handset-microphone labels, pitch track estimates, speech activity detection labels, baseline GMM-UBM acoustic scores, and word-level language model scores. The official NIST evaluation plan and lists can be found at the NIST SRE page http://www.nist.gov/speech/tests/spk/2003.

Scores from each partition are pooled and a detection error tradeoff (DET) curve is plotted to show system results at all operating points. The equal error rate (EER), where the false acceptance rate equals the missed detection rate, is used as a summary performance measure for comparing systems. Each approach formed a likelihood ratio detector by creating a speaker model using training data and a single speaker-independent background model using data from the held-out splits. For some systems, a set of individual background speaker models from the held-out set was used as cohort models. During recognition, a test utterance is scored against the speaker and background model(s) and the ratio (or difference in the log domain) is reported as the detection score for DET plotting and for fusing.

3. Features and Classifiers

In this section, we review some approaches to exploit high-level speaker information. The reader should consult the referenced papers for more details on the systems.

3.1 Spectral

The first set of features and classifiers are those based on spectral features. Three systems were applied: standard Gaussian Mixture Modeling with a Universal Background Model (GMM-UBM) system, a new Support Vector Machine (SVM) classifier, and a GMM-UBM system using only a selected subset of vocabulary words.

GMM-UBM cepstral features. [2] The basic system used is a likelihood ratio detector with target and alternative probability distributions modeled by GMM. A Universal Background Model GMM is used as the alternative hypothesis model and target models are derived using Bayesian adaptation (also known as Maximum A-Posteriori (MAP) training). Feature mapping [3] was used for channel compensation and T-norm [4] for score normalization.

SVM cepstral features. [5] The Spectral SVM system uses a novel sequence kernel that compares entire utterances using a generalized linear discriminant. The Generalized Linear Discriminant Sequence (GLDS) kernel starts with 18 LPCC and 18 delta-LPCC features vectors that are expanded into a feature space using a monomial basis. All monomials up to degree 3 were used, resulting in a feature space expansion of dimension 9139. We used a diagonal approximation to the kernel inner product matrix.

Text-constrained GMM-UBM. [6] This system is similar to the GMM-UBM baseline system but only speech from a subset of 17 words is used for all training and testing. The idea is to convert the task from text-independent to text-dependent recognition. The words were selected from a set of 80 of the most occurring words based on the minimum decision cost function value on held out data sets. The 17 words used are: (yeah, and, I, you, really, so, like, that, uh-huh, know, but, to, the, right, oh, my, just). Feature mapping and T-norm were also applied.

3.2 Prosodic

The second set of features is based on prosodic measurements, such as pitch, energy and durations. The aim here is to capture information about speaking style and cadence.

Pitch and Energy Distributions. [7] To capture the characteristic distributions of a speaker's pitch and energy values a simple GMM-UBM classifier was used with a feature vector consisting of per-frame log pitch, log energy and their first derivatives. Voice/unvoiced boundaries were respected when computing delta parameters.

Pitch and Energy Track Dynamics. [7] To model *pitch gestures* (joint pitch and energy dynamics), we converted the pitch and energy contours into a sequence of tokens reflecting the joint state of the contours (rising or falling) and then applied simple n-gram tools to model and classify distinctive token patterns from token sequences. In addition to the direction of the contour, the duration of the segment can also be integrated into the symbol sequence to provide a better characterization of the speaking style of the speaker, i.e., how long the speaker maintains certain dynamic configurations. Since we are using n-grams to model the sequence, we quantized the segment durations into 2 levels: Short and Long. Such quantization is performed separately for voice and unvoiced segments. We set the quantization levels using the mean of segment durations from held-out data. Short is assigned to voiced segments with duration less than 8 frames, and for unvoiced segments with less than 14 frames. Thus each segment symbol is now augmented with an additional duration tag: S and L, depending on if it is less than or more than a certain number of frames in duration, respectively. Additionally phone and word context can be added to these measures but was not used in this evaluation.

3.3 Phonetic

This set of features is focused on capturing speaker information carried at the phonetic level, primarily pronunciation characteristics.

GMM state N-grams [8] In this approach the sequence of GMM states is used to characterize the sub-phonetic patterns of a speaker. To do this speech is passed through a GMM tokenizer that produces a stream of symbols corresponding to the frame-by-frame indices of the highest scoring GMM component. A speaker is then modeled using a simple unconditioned (joint) n-gram model. A background model is also created using a set of held out speakers. During recognition, a likelihood ratio test between the speaker and background model for an input sequence is applied.

Phone N-grams. [9] In this approach, the time sequence of phones coming from a bank of open-loop phone recognizers is used to capture some information about speaker-dependent pronunciations. Multiple phone streams are scored independently and fused at the score level. Again, n-gram models and a likelihood-ratio classifier are used.

Phone SVM. [10] In this new discriminative system, a kernel for comparing conversation sides based upon methods from information retrieval is applied. Sequences of phones are converted to a vector of probabilities of occurrences of terms and co-occurrences of terms (bag of unigram and bag of bigrams). A weighting based upon a linearization of likelihoods is then used to compare vectors for SVM training. A background for the SVM consisted utterances taken from speakers not in the current split.

Pronunciation Modeling. [11] The aim here is to learn speaker-dependent pronunciations by comparing constrained word-level automatic speech recognition (ASR) phoneme streams with open-loop phone streams. The phonemes from the CMU Sphinx 3.3 ASR word transcripts were aligned on a per-frame level with open-loop phoneme transcripts. Conditional probabilities for each open-loop phone, given an ASR phoneme, are computed per speaker and for a background model. A likelihood ratio test between the two models is applied in testing.

3.4 Idiolectal

The focus here is to capture high-level information about the word usage (idiolect) of a particular speaker. This is the speech analog to various methods of author identification, where writers are characterized by their written texts.

Word N-grams. [12] In this approach, unconditioned n-gram models of word transcripts from an ASR system of several conversations from a speaker are used to model the speaker' idiolectal patterns. A background idiolect model from a large population of held-out speakers is used to characterize general idiolect patterns, and a likelihood ratio text between the two models is used for testing. It is particularly interesting that reasonable performance can be obtained even using a highly errorful transcript (approx 50% word error rate). T-norm was also applied using speakers from held-out sets.

3.5 Dialogic

The aim with these features is to capture long-term interaction patterns that can help characterize a speaker.

Conversational Pattern N-grams. N-grams from conversational patterns, an additional level of linguistic information, are extracted from the transcripts for training and testing. Intuitively, we know that different speakers behave differently in conversation. Some people tend to dominate conversations; others work to get a word in edge-wise. Speakers may take turns dominating a conversation. This system is based on a simple but novel n-gram notation that is intended to pick up on these kinds of behavioral patterns in conversation. The conversational patterns are represented with a very simple notation that indicates the duration and amount of text content in the speech transcript for each utterance. Although we only have the transcript for the test/target speakers, we are able to infer the duration of the other speakers' part of the conversation. The simplest form of notation we tried was to just mark the duration of the turns. Additionally, for the test/target speakers we also assigned labels for the amount of text in the transcript (a crude measure of speaker's "baud rate"). To construct the label, we use the labels that represent duration, number of bytes, and number of words.

3.6 Semantic

The last level of information, semantic, is the more specialized one and not pursued for the EDT. Given time-specific, world knowledge of a person's current interests or needs, one could construct a classifier looking for particular topic-related words or phrases to use in conjunction with the other classifiers. For example, from previous emails, a call-center may know that a particular person is having trouble with billing for phone service, and so a call inquiring about the topic of billing problems would more likely be from that particular person.

3.7 Fusion

The scores from the systems were fused with a perceptron classifier using *LNKnet* [13]. The perceptron architecture chosen has N input nodes (where N is the number of systems being fused), no hidden layers, and two output nodes (target and nontarget). Input values to the perceptron were normalized to zero mean and unit standard deviation using parameters derived from the training data. The perceptron weights were trained with 10-fold jacknife for each of the training conversation sets {1, 2, 4, 8, and 16}. The classifier corresponding to the number of training conversations is then used to fuse scores from systems. A more detailed description of the fusion system and other experiments on the NIST EDT 2001 data can be found in [14].

4. Results on 2003 EDT

In this section we present results of applying the above systems to the 2003 EDT. A list of the systems used in for system fusion experiments is shown in Table 1. These systems were selected to span the different levels of information. For analysis purposes, we will group the above systems into spectral and non-spectral based systems. The spectral based systems are systems 0, 1 and 2. The non-spectral are systems 3-9.

Table 1: Systems used in fusion experiments

System Number	Component System Descriptions
0	GMM-UBM Baseline
1	Text-Constrained GMM-UBM
2	LPCC SVM
3	Phone SVM
4	Word n-gram (baseline idiolect)
4t	Word n-gram with T-norm
5	Phone n-gram
6	Pitch & Energy GMM
7	Slope & Duration n-gram
8	Pronunciation
9	Conversational Patterns n-gram

4.1 System Combination Results

In Figure 2 we show the equal error rate (EER) as a function of number of training conversations for the three individual spectral systems as well as the fusion of the three. The best single system is the SVM but we see a significant gain in fusing all three systems. This complementary fusion of the generative and discriminative systems was also observed in the standard NIST speaker detection evaluation. While all systems improve with increasing number of training conversations, the text-constrained GMM-UBM system benefits the most. It is likely that the spectral systems continue to improve with the number of training conversations due to increased session and channel variability in the training data rather than the increase in amount of training data.

In Figure 3, we show the EER as a function of number of training conversations for the non-spectral systems. Note that T-normed results are not shown here. The best single performing system is the phone SVM followed by almost identical performance for the phone n-gram and pronunciation systems. We believe the pronunciation system was hampered by some poor phoneme alignments. The word n-gram and conversational patterns n-gram systems have the largest gains with increasing number of training conversations, which is not unexpected since they rely on using events that are not as frequently occurring as other features. We also see again that the fusion of these different levels of information produces a gain in performance over the individual systems.

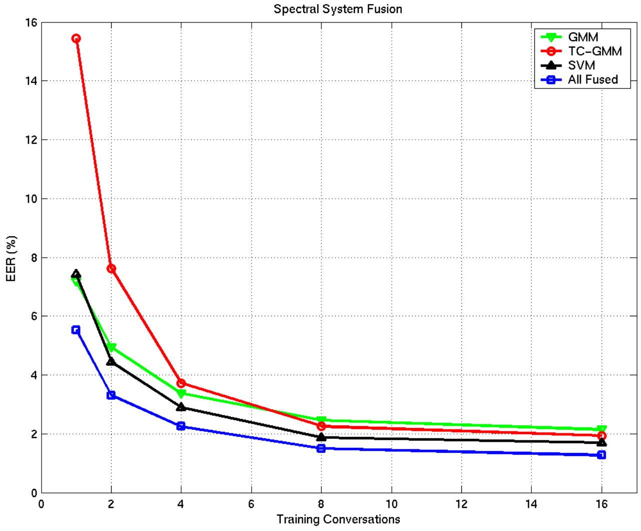


Figure 2: EER vs. number of training conversations for spectral systems.

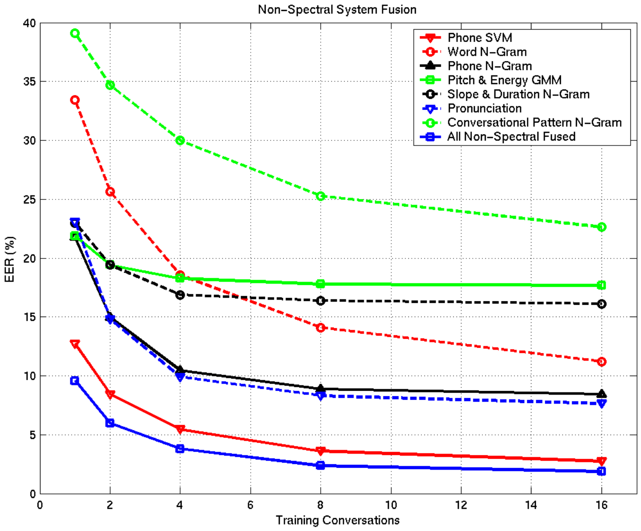


Figure 3: EER vs. number of training conversations for non-spectral systems.

Finally, in Figure 4, we show the EER versus the number of training conversations for the spectral, non-spectral, fusion of all and minimum DCF search ‘oracle’ system. The oracle system is an exhaustive search over all system combinations to find the set that minimizes the DCF value and is meant for diagnostics purposes to see which systems contribute the most.

The spectral systems outperform the non-spectral systems, but the gap is relatively small for 8-16 conversation cases. We also see, as was the case from the SuperSID workshop, that the combination of spectral and non-spectral systems improves the overall error rate. The gain is not as great as observed with Switchboard-I data [1], but this is most likely due to more handset mismatch conditions in the Switchboard-II data[ii].

[ii] In Switchboard-II, callers were required to use a different phone number when placing an incoming call, thus presumably increasing the handset variability of the data. Switchboard-I did not have this requirement and so had less handset variability.

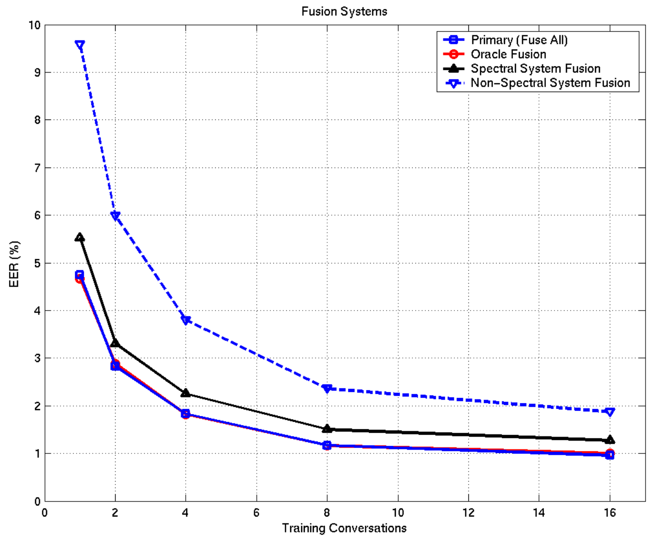


Figure 4: EER vs. number of training conversations for fusion of spectral and non-spectral systems.

The complete Detection Error Tradeoff (DET) curves for the all-system fusion are shown in Figure 5. A post-evaluation experiment determined that we could reduce the EER for 8-conversation training to < 1%.

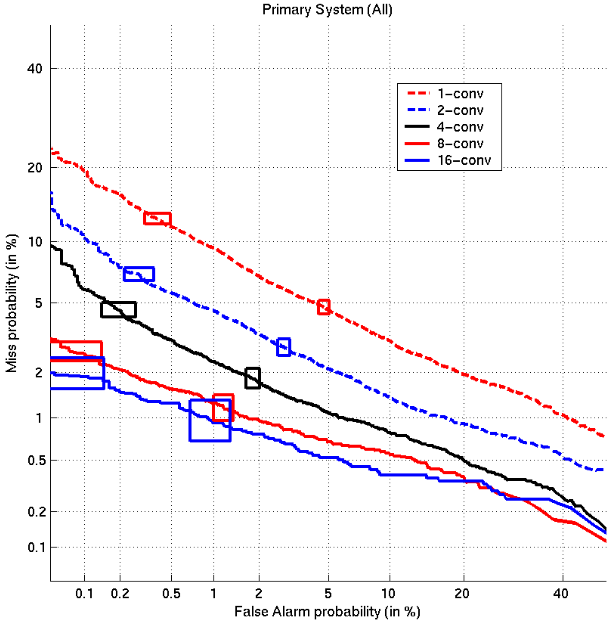


Figure 5: DET curves for the all-system fusion. The boxes represent 95% confidence regions for the EER and min DCF operating points.

4.2 Matched and Unmatched Analysis

It is well known that better speaker recognition performance is expected when the training and testing data come from a common handset (matched conditions). While we do not have explicit handset information, we do have coded versions of the telephone number used for each call, from which we can assume that a different phone number implies a different handset.

For the 8 conversation training condition, we defined a matched target trial as the case when the telephone number of the test conversation matches any one of the telephone numbers used by the target in his/her training conversations. An unmatched trial is when there is no overlap between train and test telephone numbers. Generally, all non-target trials will be unmatched trails[iii]. For the 8 conversation training condition, approximately 50% of the target trials were labeled as matched conditions.

In Figure 6 we plot DET curves from the 8-conversation all-system fusion for matched, unmatched and all target trial cases. The non-target trials are constant for each curve. For the matched target trials the EER is 0.6%, compared to 1.5% for the unmatched target trials and 1.2% for all target trials.

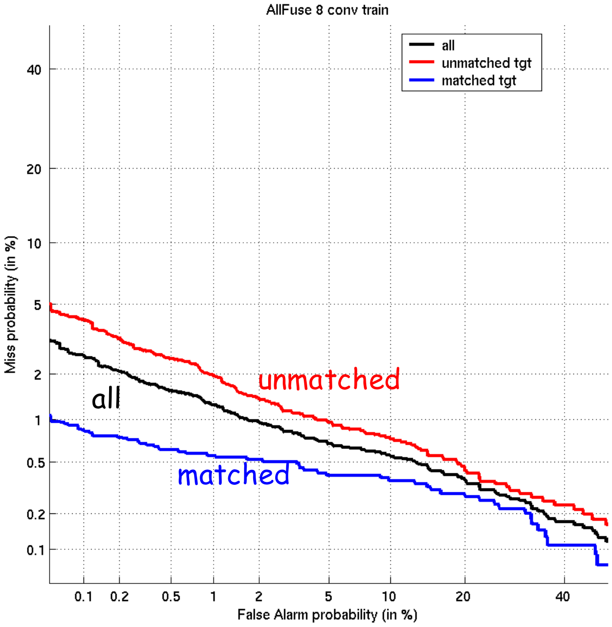


Figure 6: DET curves for the 8-conversation train all-system fusion. Curves for matched and unmatched target trial conditions are shown.

A break down of EER for matched and unmatched cases from the spectral, non-spectral and all-system fusion systems is shown in Figure 7. Here we see that all systems have a loss in performance under the unmatched case. We also see, however, that the fusion of spectral and non-spectral systems shows improved performance under

[iii] This does not occur when speakers share a common phone number.

the matched and unmatched cases. Additionally, the relative loss in ERR going from matched to unmatched for the non-spectral system is slightly better than that of the spectral systems. The non-spectral system was not as robust to the handset mismatch as was hoped. This is not totally unexpected, since the non-spectral features (pitch, phones, words) are derived at some point directly from the acoustic waveform and so are subject to the biases induced by varying handsets. Current work is focused on better understanding these biases and examining ways to mitigate their effects in the non-spectral features.

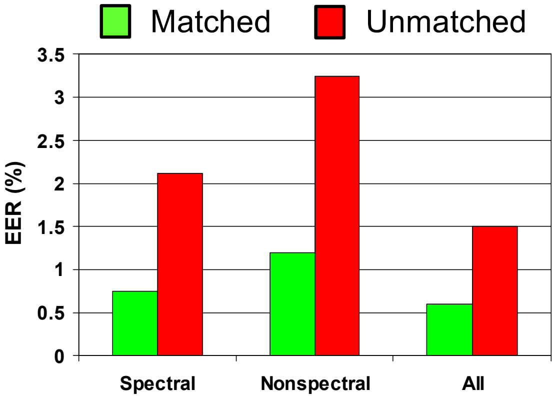


Figure 7: EER from spectral, non-spectral and all system for matched and unmatched cases.

5. Conclusions and Acknowledgements

In this paper, we have outlined some of the recent trends and systems aimed at moving beyond low-level, short-term spectra by exploiting high-level information for speaker recognition. These systems focus on capturing speaker habits and idiosyncrasies as manifest in different aspects of speech. Even at low error rates, it was shown that there is still significant benefit in combining complementary types of information. An initial analysis of the data for matched and unmatched target trials, shows there is still significant loss under mismatched conditions, but fusion of different levels of information still is beneficial. Further work is aimed at a more detailed error analysis to better understand under what conditions different information sources best help performance. The aim would be to learn how to better combine systems.

The authors wish to thank Andre Adami of OGI for running the prosodic experiments and David Klusácek of Charles University for running the pronunciations experiments.

6. References

[1] D. A. Reynolds, W. D. Andrews, J. P. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Klusácek, J. S. Abramson, R. Mihaescu, J. J. Godfrey, D. A. Jones and B. Xiang, *The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Hong Kong, 2003, pp. 784-787.

[2] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Mixture Models", Digital Signal Processing, Vol. 10, pp. 181-202, 2000

[3] D. A. Reynolds, *Channel Robust Speaker Verification via Feature Mapping,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Hong Kong, 2003, pp. 53-56.

[4] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, *Score Normalization for Text-Independent Speaker Verification Systems,* Digital Signal Processing, 10 (2000), pp. 42-54.

[5] W. M. Campbell, *A SVM/HMM System for Speaker Recognition,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Hong Kong, 2003, pp. 209-302.

[6] D. E. Sturim, D. A. Reynolds, R. B. Dunn and T. F. Quatieri, *Speaker Verification using Text-Constrained Gaussian Mixture Models,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Orlando, Florida, 2002.

[7] A. Adami, R. Mihaescu, D. A. Reynolds and J. J. Godfrey, *Modeling Prosodic Dynamics for Speaker Recognition,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Hong Kong, 2003, pp. 788-791.

[8] B. Xiang, *Text-Independent Speaker Verification with Dynamic Trajectory Model,* IEEE Signal Processing Letters, 10 (2003), pp. 141-143.

[9] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey and J. Hernández-Cordero, *Gender-Dependent Phonetic Refraction for Speaker Recognition,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Orlando, Florida, 2002, pp. 149-152.

[10] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jone, T. R. Leek, *Phonetic Speaker Recognition with Support Vector Machines*, to appear Neural Information Processing Systems (NIPS) Conference, 2003.

[11] D. Klusácek, J. Navrátil, D. A. Reynolds and J. P. Campbell, *Conditional Pronunciation Modeling in Speaker Detection,* International Conference on Acoustics, Speech, and Signal Processing, IEEE, Hong Kong, 2003, pp. 804-807.

[12] G. Doddington, *Speaker Recognition based on Idiolectal Differences between Speakers,* Eurospeech, ISCA, Aalborg, Denmark, 2001, pp. 2517-2520.

[13] R. P. Lippmann, L. C. Kukolich and E. Singer, *LNKnet: Neural Network, Machine-Learning, and Statistical Software for Pattern Classification,* Lincoln Laboratory Journal, 6 (1993), pp. 249-268.

[14] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, *Fusing High- and Low-Level Features for Speaker Recognition,* Eurospeech ISCA, Geneva Switzerland, 2003.

Invited Speaker

James Wayman
San Jose State University

Biometric Testing

Abstract

Large-scale biometric testing has a history of at least 25 years. However, each test has used different testing and reporting protocols, making results very hard to understand and causing longitudinal comparison of biometric device and system performance to be impossible. Working Group 5 of the ISO/IEC JTC1 Standing Committee 37 on biometrics has been established to address these variances in the hope that a single test and reporting standard can be developed. But many within the field feel that a single "standard" might not be possible. In this talk, we will review historical testing and reporting protocols, point out the areas of controversy, and analyze in detail the recent UK contribution to the SC37 WG5 process.

Author Index

Notes

Notes

Notes

Notes

Notes