Combining Biometric Evidence for Person Authentication

J. Bigun[1], J. Fierrez-Aguilar[2]*, J. Ortega-Garcia[2], and J. Gonzalez-Rodriguez[2]

[1] Halmstad University, Sweden
josef.bigun@ide.hh.se
[2] Universidad Politecnica de Madrid, Spain
{jfierrez,jortega,jgonzalez}@diac.upm.es

Abstract. Humans are excellent experts in person recognition and yet they do not perform excessively well in recognizing others only based on one modality such as single facial image. Experimental evidence of this fact is reported concluding that even human authentication relies on multimodal signal analysis. The elements of automatic multimodal authentication along with system models are then presented. These include the machine experts as well as machine supervisors. In particular, fingerprint and speech based systems will serve as illustration. A signal adaptive supervisor based on the input biometric signal quality is evaluated. Experimental results on data collected from a mobile telephone prototype application are reported demonstrating the benefits of the reported scheme.

1 Introduction

Face recognition is an important element of person authentication in humans. Human face analysis engages special signal processing in visual cortex different than processing of other objects [2, 3]. It is reliably observed in a number of studies that negative bias in ability to recognize faces of another racial group versus own racial group exists [4, 5, 6]. It has been confirmed that the hair style and facial expressions are significant distraction factors for humans. It has recently been revealed [7] that the lack of caricature type information hampers the recognition more than the lack of silhouette and shading information and that there is a gender bias in women's and and men's abilities to recognize faces. In [7] it is shown that, depending on the gender to be recognized, humans were able to recognize unfamiliar faces from photographs at the success rate of 55-75%. This suggests that multimodal biometric information processing e.g. using signals from body motion including the head motion, speech, and lip movements, plays a significant role in human's efforts of authenticating other humans.

Automatic access of persons to services is becoming increasingly important in the information era. Although person authentication by machine has been a

* This study has been carried out while J. F.-A. and J. O.-G. were guest scientists at Halmstad University [1].

M. Tistarelli, J. Bigun, and E. Grosso (Eds.): Biometrics School 2003, LNCS 3161, pp. 1–18, 2005.

subject of study for more than thirty years [8, 9], it has not been until recently that the matter of combining a number of different traits for person verification has been considered [10, 11]. There are a number of benefits of doing so, just to name a few: false acceptance and false rejection error rates decrease, the authentication system becomes more robust against individual sensor or subsystem failures and the number of cases where the system is not able to give an answer (e.g. bad quality fingerprints due to manual work or larynx disorders) vanishes. The technological environment is also appropriate because of the widespread deployment of multimedia-enabled mobile devices (PDAs, 3G mobile phones, tablet PCs, laptops on wireless LANs, etc.). As a result, much research work is currently being done in order to fulfill the requirements of applications for masses.

Two early sound theoretical frameworks for combining different machine experts in a multimodal biometric system are described in [11] and [12], the former from a risk analysis perspective [13] and the later from a statistical pattern recognition point of view [14]. Both of them concluded (under some mild conditions which normally hold in practice) that the weighted average is a good way of conciliating the different experts. Soon after, multimodal fusion was studied as a two-class classification problem by using a number of machine learning paradigms [15, 16, 17], for example: neural networks, decision trees and support vector machines. They too confirmed the benefits of performance gains with trained classifiers, and favored support vector machines over neural networks and decision trees. The architecture of the system, ease of training, ease of implementation and generalization to mass use were however not considered in these studies. As happens in every pattern recognition problem which is application-oriented, these are important issues that influence the choice of a supervisor.

Interestingly enough, some recent works have nevertheless reported comparable performance between fixed and trained combining strategies [18, 19] and a debate has come out investigating the benefits of both approaches [20, 21]. As an example, and within this debate, some researches have shown how to learn user-specific parameters in a trained fusion scheme [22, 23]. As a result, they have showed that the overall verification performance can be improved significantly by considering user-dependent fusion schemes.

In this work we focus on some other benefits of a trained fusion strategy. In particular, an adaptive trained fusion scheme is introduced here. With adaptive fusion scheme, we mean that the supervisor readapts to each identity claim as a function of the quality of the input biometric signal, usually depending on external conditions such as light and background noise. Furthermore, experiments on real data from a prototype mobile authentication application combining fingerprint and speech data are reported.

This paper is structured as follows. In Section 2, we summarize the findings on mono-modal human recognition performance suggesting that individual modalities do not have to score high to yield robust multimodal systems [7]. Beginning in section 3 with some definitions, we discuss machine supervisors for multimodal authentication [1, 11] in the sequel. The elements of multimodal authentication along with major notations are introduced in section 4. In section 5, the statistical framework for conciliating the different expert opinions together with simplified and full supervisor algorithms are described. The components of our prototype mobile authentication application, namely fingerprint and speaker verification subsystems, are briefly described in section 6. Some experiments are reported in section 8 using the above mentioned multimodal authentication prototype and the performance evaluation methodology described in section 7. Conclusions will be finally given in section 9.

2 Human Face Recognition Performance

There is a general agreement on that, approximately at the age of 12 the performance of children in face recognition reaches adult levels, that there is already an impressive face recognition ability by the age of 5 and that measurable preferences for face stimuli exist in babies even younger than 10 minutes [24].

Our study [7], that aimed at quantifying the skills of humans in face recognition of unfamiliar faces, has been supported by more than 4000 volunteers[3]. We found that the lack of high spatial frequencies in visual stimuli, which result in blurred images as if face information were coming from an unfocused camera, hamper the recognition significantly more than the lack of low spatial frequencies, which result in stimuli similar to artist drawn faces, see Figure 1.

The face recognition questions. In all 8 questions (Q1-Q8) the task was to identify the picture of a stimulus person among a query set consisting of 10 pictures. The subjects were informed, before the start of the test, that the stimulus image and the image to be found in the query set were taken at two different occasions and that these two images could differ significantly in hair style, glasses, expression of the face, facial hair, clothing, etc. due to the natural changes in appearance that occur upon passage of time (a few months). In Q1-Q4 and Q8 the stimulus and the query set were shown simultaneously, in the same screen. Questions Q5-Q7 were similar to the other questions except that they included a memorization task in that the stimulus was shown in its own page without the query set. When the subject wished to continue, the stimulus was replaced by the query set, forcing the subjects to answer the question without a possibility to see the stimulus.

The available results [7] reveal that in questions in which the face image to be recognized was not manipulated (e.g. the high frequencies were not depreciated), the recognition rate varied between 55-75 % in the average. A surprising result was that the females had in the average better success in all tasks than the males. A typical question in the test is illustrated by Figure 1.

The fact that the success rates are in the best cases (female subjects) around 80% suggests that not only mono-modal information but also multimodal biometric information processing e.g. using the signals from body motion including head motion, speech, and lip movements, plays a significant role when humans authenticate other humans.

[3] As of November 2003. The test is available at http://www.hh.se/facetest

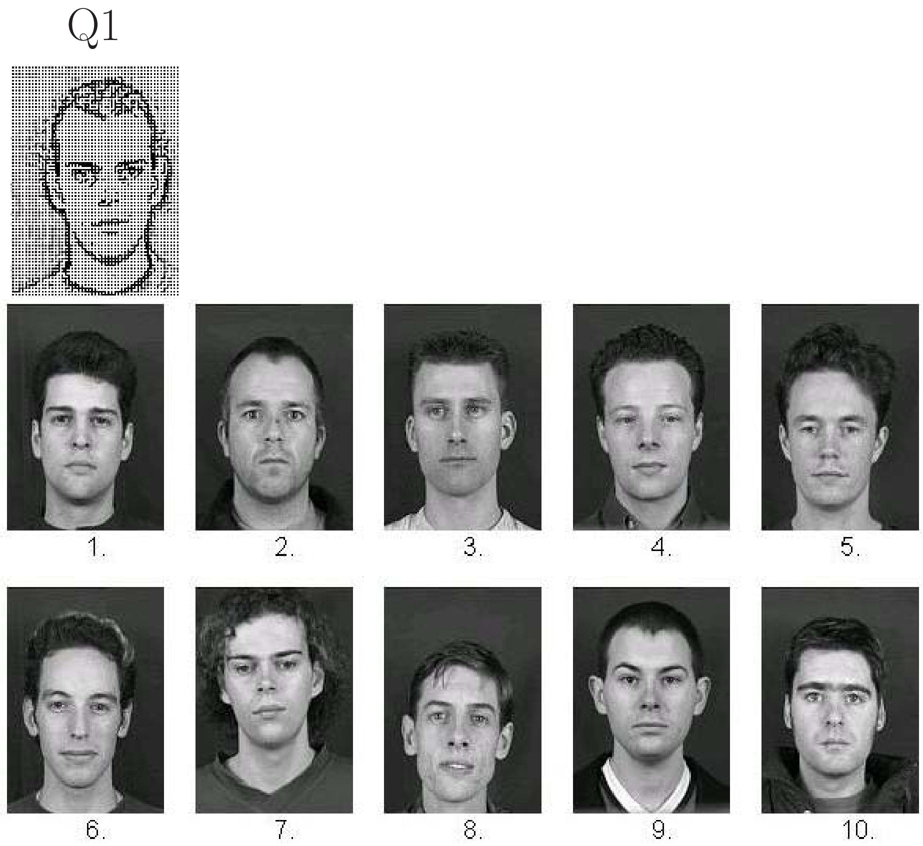

Fig. 1. A question illustrating the test. On the top the stimulus is given. The subject matched the stimulus with one of the 10 images below the stimulus. The low spatial frequencies of the stimulus were removed by signal processing.

3 Definitions

In *authentication* (also known as *verification*) applications, the users or *clients* are known to the system whereas the *impostors* can potentially be the world population. In such applications the users provide their claimed identities (either true or false) and a one-to-one matching is performed. If the result of the comparison (also *score* or *opinion*) is higher than a *verification threshold*, then the claim is accepted, otherwise the claim is rejected.

In *identification* applications, there is no identity claim and the candidate is compared to a database of client models, therefore a one-to-many matching is performed in this case. In the simplest form of identification, also known as *closed-set identification*, the best client model is selected. In *open-set identification*, the highest score is further compared to a verification threshold so as

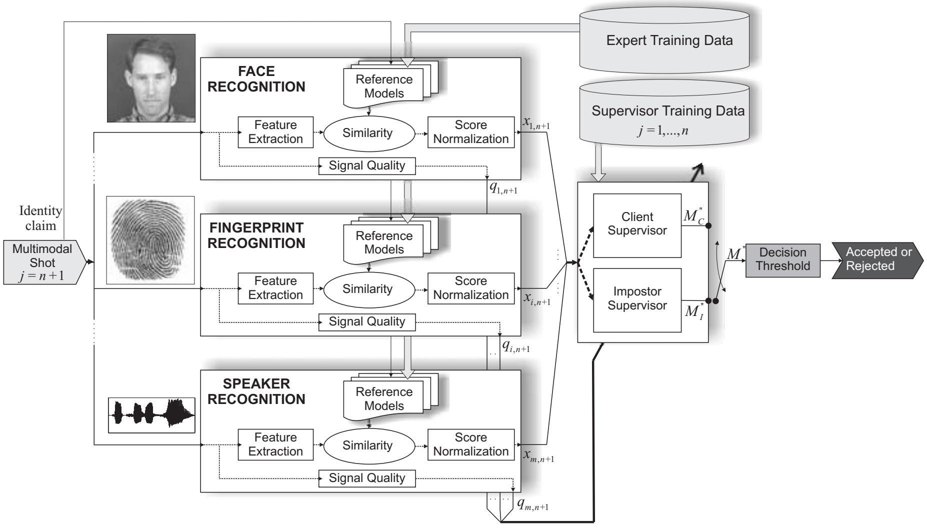

Fig. 2. The proposed system model of multi-modal person authentication.

to accept/reject this candidate as belonging or not to the database (an implicit authentication step).

In a multimodal authentication framework, various subsystems (also denoted as *experts*) are present, each one of them specialized on a different trait. Each expert delivers its opinion on a "package" of data containing an identity claim (e.g. face images, fingerprint images, speech data, etc.) that will be referred to as a *shot*. This paper is focused on combining the experts opinions (also known as *soft decisions*). It will be shown that a careful design of the *supervisor* (also known as *fusion* strategy) yields a combined opinion which is more reliable than the best expert opinion.

4 System Model

Below is a list of the major notations we use throughout the paper, see also Figure 2.

i Index of the experts, $i \in 1 \ldots m$
j Index of the shots, $j \in 1 \ldots n, n+1$
x_{ij} Authenticity score delivered by expert i on shot j
s_{ij} Variance of x_{ij} as estimated by expert i
y_j The true authenticity score of shot j
z_{ij} The error score of an expert $z_{ij} = y_j - x_{ij}$

Note that the experts are allowed to provide a quality of the score which is modelled to be inversely proportional to s_{ij}. This strategy is novel with respect to the implemented supervisors reported so far in that it is the expert who is providing a variance on every authenticity score it delivers, not the supervisor. It is also worth pointing out that y_j can take only two numerical values corresponding to "False" and "True". If x_{ij} is between 0 and 1 then these values are chosen to be 0 and 1 respectively. We assume that the experts have been trained on other shots apart from $j \in 1 \ldots n, n+1$. The supervisor is trained on shots $j \in 1 \ldots n$ (i.e. x_{ij} and y_j are known for $j \in 1 \ldots n$) and we consider shot $n+1$ as a test shot on the working multimodal system (i.e. $x_{i,n+1}$ is known, but y_{n+1} is not known and the supervisor task is to estimate it).

5 Statistical Model

The model for combining the different experts is based on Bayesian statistics and the assumption of normal distributed expert errors, i.e. z_{ij} is considered to be a sample of the random variable $Z_{ij} \sim N(b_i, \sigma_{ij}^2)$. It has been shown experimentally [11] that this assumption does not strictly hold for common audio- and video-based biometric machine experts, but it is shown that it holds reasonably well when client and impostor distributions are considered separately. Taking this result into account, two different supervisors are constructed, one of them based on expert opinions where $y_j = 1$

$$\mathcal{C} = \{x_{ij}, s_{ij} | y_j = 1 \text{ and } 1 \leq j \leq n\} \tag{1}$$

while the other is based on expert opinions where $y_j = 0$

$$\mathcal{I} = \{x_{ij}, s_{ij} | y_j = 0 \text{ and } 1 \leq j \leq n\} \tag{2}$$

The two supervisors will be referred to as *client supervisor* and *impostor supervisor*, respectively (see Figure 2).

The client supervisor estimates the expected true authenticity score of an input claim based on its expertise on recognizing client data. More formally, it computes $M''_{\mathcal{C}} = E\left[Y_{n+1} | \mathcal{C}, x_{i,n+1}\right]$ (the prime notation will become apparent later on). In case of impostor supervisor, $M''_{\mathcal{I}} = E\left[Y_{n+1} | \mathcal{I}, x_{i,n+1}\right]$ is computed. The conciliated overall score M'' takes into account the different expertise of the two supervisors and chooses the one which came closest to its goal, i.e. 0 for the impostor supervisor and 1 for the client supervisor:

$$M'' = \begin{cases} M''_{\mathcal{C}} \text{ if } |1 - M''_{\mathcal{C}}| - |0 - M''_{\mathcal{I}}| < 0 \\ M''_{\mathcal{I}} \text{ otherwise} \end{cases} \tag{3}$$

Based on the normality assumption of the errors, the supervisor algorithm described in [11] is obtained (see [13] for further background and details). In the following, we summarize this algorithm in the two cases where it can be applied.

5.1 Simplified Supervisor Algorithm

When no quality information is available, the following simplified supervisor algorithm is obtained by using $s_{ij} = 1$:

1. (Supervisor Training) Estimate the bias parameters of each expert. In case of the client supervisor the bias parameters are

$$M_{\mathcal{C}i} = \frac{1}{n_{\mathcal{C}}} \sum_j z_{ij} \quad \text{and} \quad V_{\mathcal{C}i} = \frac{\alpha_{\mathcal{C}i}}{n_{\mathcal{C}}} \tag{4}$$

   where j indexes the training set $\mathcal{C}$, $n_{\mathcal{C}}$ is the number of shots in $\mathcal{C}$ and

$$\alpha_{\mathcal{C}i} = \frac{1}{n_{\mathcal{C}} - 3} \left(\sum_j z_{ij}^2 - \frac{1}{n_{\mathcal{C}}} \left(\sum_j z_{ij} \right)^2 \right) \tag{5}$$

   Similarly $M_{\mathcal{I}i}$ and $V_{\mathcal{I}i}$ are obtained for the impostor supervisor.
2. (Authentication Phase) At this step, both supervisors are operational, so that the time instant is always $n+1$ and the supervisors have access to expert opinions $x_{i,n+1}$ but not access to the true authenticity score y_{n+1}. Client and impostor supervisors calibrate the experts according to their past performance, yielding (for the client supervisor)

$$M'_{\mathcal{C}i} = x_{i,n+1} + M_{\mathcal{C}i} \quad \text{and} \quad V'_{\mathcal{C}i} = (n_{\mathcal{C}} + 1) V_{\mathcal{C}i} \tag{6}$$

   and then the different calibrated experts are combined according to

$$M''_{\mathcal{C}} = \frac{\sum_{i=1}^{m} \frac{M'_{\mathcal{C}i}}{V'_{\mathcal{C}i}}}{\sum_{i=1}^{m} \frac{1}{V'_{\mathcal{C}i}}} \tag{7}$$

   Similarly, $M'_{\mathcal{I}i}$, $V'_{\mathcal{I}i}$ and $M''_{\mathcal{I}}$ are obtained. The final supervisor opinion is obtained according to (3).

The algorithm described above has been successfully applied in [25] in a multimodal authentication system combining face and speech data. Verification performance improvements of almost an order magnitude were reported as compared to the best modality.

5.2 Full Supervisor Algorithm

When not only the experts scores but also the quality of the scores are available, the following algorithm is obtained:

1. (Supervisor Training) Estimate the bias parameters. For the client supervisor

$$M_{\mathcal{C}i} = \frac{\sum_j \frac{z_{ij}}{\sigma_{ij}^2}}{\sum_j \frac{1}{\sigma_{ij}^2}} \quad \text{and} \quad V_{\mathcal{C}i} = \frac{1}{\sum_j \frac{1}{\sigma_{ij}^2}} \tag{8}$$

where the training set $\mathcal{C}$ is used. The variances σ_{ij}^2 are estimated through $\bar{\sigma}_{ij}^2 = s_{ij} \cdot \alpha_{\mathcal{C}i}$, where

$$\alpha_{\mathcal{C}i} = \frac{1}{n_{\mathcal{C}} - 3} \left(\sum_j \frac{z_{ij}^2}{s_{ij}} - \left(\sum_j \frac{z_{ij}}{s_{ij}} \right)^2 \left(\sum_j \frac{1}{s_{ij}} \right)^{-1} \right) \tag{9}$$

Similarly $M_{\mathcal{I}i}$ and $V_{\mathcal{I}i}$ are obtained for the impostor supervisor.

2. (Authentication Phase) First the supervisors calibrate the experts according to their past performance, for the client supervisor

$$M'_{\mathcal{C}i} = x_{i,n+1} + M_{\mathcal{C}i} \quad \text{and} \quad V'_{\mathcal{C}i} = s_{i,n+1}\alpha_{\mathcal{C}i} + V_{\mathcal{C}i} \tag{10}$$

and then the different calibrated experts are combined according to (7). Similarly, $M'_{\mathcal{I}i}$, $V'_{\mathcal{I}i}$ and $M''_{\mathcal{I}}$ are obtained. The final supervisor opinion is obtained according to (3).

The algorithm described above has been successfully applied in [13] combining human expert opinions but not in a multimodal authentication application.

5.3 Adaptive Strategy

The variance s_{ij} of the score x_{ij} is provided by the expert and concerns a particular authentication assessment. It is not a general reliability measure for the expert itself, but a certainty measure based on qualitative knowledge of the expert and the data the expert assesses. Typically the variance of the score is chosen as the width of the range in which one can place the score. Because such intervals can be conveniently provided by a human expert, the algorithm in section 5.2 constitutes a systematic way of combining human and machine expertise in an authentication application. An example of such an application is forensics, where machine expert approaches have been proposed [26] and human opinions must be taken into consideration.

In this work, we propose to calculate s_{ij} for a machine expert by using a quality measure of the input biometric signal (see Figure 2). This implies taking into account equation (10) right, that the trained supervisor adapts the weights of the experts using the input signal quality. First we define the quality q_{ij} of the score x_{ij} according to

$$q_{ij} = \sqrt{Q_{ij} \cdot Q_{i,claim}} \tag{11}$$

where Q_{ij} and $Q_{i,claim}$ are respectively the quality label of the biometric sample used by expert i in shot j and the average quality of the biometric samples used by expert i for modelling the claimed identity. The two quality labels Q_{ij} and $Q_{i,claim}$ are supposed to be in the range $[0, q_{max}]$ with $q_{max} > 1$ where 0 corresponds to the poorest quality, 1 corresponds to normal quality and q_{max} corresponds to the highest quality. Finally, the variance parameter is calculated according to

$$s_{ij} = \frac{1}{q_{ij}^2} \tag{12}$$

6 Monomodal Experts

6.1 Speaker Expert

For the experiments reported in this work, the GMM-based speaker expert from Universidad Politecnica de Madrid used in the 2002 NIST Speaker Recognition evaluation [27] has been used. Below we briefly describe the basics, for more details we refer to [27, 28].

Feature extraction. Short-time analysis of the speech signal is carried out by using 20 ms Hamming windows shifted 10 ms. For each analysis window $t \in [1, 2, \ldots, T]$, a feature vector $\mathbf{m}_t$ based on Mel-Frequency Cepstral Coefficients (MFCC) and including first and second order time derivative approximations is generated. The feature vectors $M = \{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_T\}$ are supposed to be drawn from a user-dependent Gaussian Mixture Model λ which is estimated in the enrollment phase via MAP adaptation of a Universal Background Model λ_{UBM}. For our tests, the UBM is a text-independent 128 mixture GMM which was trained by using approximately 8 hours of Spanish mobile speech data (gender balanced).

Similarity computation. Given a test utterance parameterized as M and a claimed identity modeled as λ, a matching score x'_{ij} is calculated by using the log-likelihood ratio

$$x'_{ij} = \log\left(p\left[M|\lambda\right]\right) - \log\left(p\left[M|\lambda_{UBM}\right]\right) \tag{13}$$

Score normalization. In order to generate an expert opinion x_{ij} between 0 and 1, the matching score x'_{ij} is further normalized according to

$$x_{ij} = \frac{1}{1 + e^{-c \cdot x'_{ij}}} \tag{14}$$

The parameter c has been chosen heuristically on mobile speech data not used for the experiments reported here.

6.2 Fingerprint Expert

For the experiments reported in this work, the minutiae-based fingerprint expert described in [29] has been used. Below we describe the basics, for more details we refer to [29, 30].

Image enhancement. The fingerprint ridge structure is reconstructed according to: *i*) grayscale level normalization, *ii*) orientation field calculation, according to [31] *iii*) interest region extraction, *iv*) spatial-variant filtering according to the estimated orientation field, *v*) binarization, and *vi*) ridge profiling.

Feature extraction. The minutiae pattern is obtained from the binarized profiled image as follows: *i*) thinning, *ii*) removal of structure imperfections from the thinned image, and *iii*) minutiae extraction. For each detected minutia, the following parameters are stored: *a*) the x and y coordinates of the minutia, *b*) the orientation angle of the ridge containing the minutia, and *c*) the x and y coordinates of 10 samples of the ridge segment containing the minutia. An example fingerprint image from MCYT Database [32], the resulting binary image after image enhancement, the detected minutiae superimposed on the thinned image and the resulting minutiae pattern are shown in Figure 3.

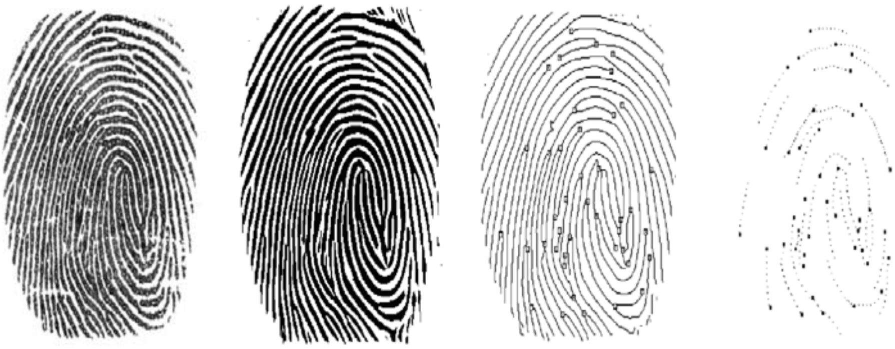

Fig. 3. Fingerprint feature extraction process

Similarity computation. Given a test and a reference minutiae pattern, a matching score x'_{ij} is computed. First, both patterns are aligned based on the minutia whose associated sampled ridge is most similar. The matching score is computed then by using a variant of the edit distance on polar coordinates and based on a size-adaptive tolerance box. When more than one reference minutiae pattern per client model are considered, the maximum matching score obtained by comparing the test and each reference pattern is used.

Score normalization. In order to generate an expert opinion x_{ij} between 0 and 1, the matching score x'_{ij} is further normalized according to

$$x_{ij} = \tanh\left(c \cdot x'_{ij}\right) \tag{15}$$

The parameter c has been chosen heuristically on fingerprint data not used for the experiments reported here.

7 Verification Performance Evaluation

Biometric verification can be considered as a detection task, involving a tradeoff between two type of errors: *i*) Type I error, also denoted as *False Rejection*

(FR) or miss (detection), occurring when a client, target, genuine, or authorized user is rejected by the system, and *ii*) Type II error, known as *False Acceptance* (FA) or false alarm, taking place when an unauthorized or impostor user is accepted as being a true user. Although each type of error can be computed for a given decision threshold, a single performance level is inadequate to represent the full capabilities of the system and, as such a system has many possible operating points, it is best represented by a complete performance curve. These total performance capabilities have been traditionally shown in form of ROC (Receiver -or also Relative- Operating Characteristic) plots, in which FA rate versus FR rate is depicted. A variant of this, the so-called DET (Detection Error Tradeoff) plot [33], is used here; in this case, the use of a normal deviate scale makes the comparison of competing systems easier. Moreover, the DET smoothing procedure introduced in [34], which basically consists in Gaussian Mixture Model estimation of FA and FR curves, has been also applied.

A specific point is attained when FAR and FRR coincide, the so-called EER (equal error rate); the global EER of a system can be easily detected by the intersection between the DET curve of the system and the diagonal line $y = x$. Nevertheless, and because of the step-like nature of FAR and FRR plots, EER calculation may be ambiguous according to the above-mentioned definition, so an operational procedure for computing the EER must be followed. In the present contribution, the procedure for computing the EER proposed in [35] has been applied.

8 Experiments

8.1 Database Description and Expert Protocol

Cellular speech data consist of short utterances in Spanish (the mobile number of each user). 75 users have been acquired, each one of them providing 10 utterance samples from 10 calls (within a month interval). The first 3 utterances are used as expert training data and the other 7 samples are used as expert test data. The recordings were carried out by a dialogue-driven computer-based acquisition process, and data were not further supervised. Moreover, 10 real impostor attempts per user are used as expert testing data, where each impostor knew the true mobile number and the way it was pronounced by the user he/she was forgering. Taking into account the automatic acquisition procedure and the highly skilled nature of the impostor data, near worst-case scenario has been prevailing in our experiments.

Fingerprint data from MCYT corpus has been used. For a detailed description of the contents and the acquisition procedure of the database, see [32]. Below, some information related to the experiments we have conducted is briefly described.

MCYT fingerprint subcorpus comprises 330 individuals acquired at 4 different Spanish academic sites by using high resolution capacitive and optical capture devices. For each user, the 10 prints were acquired under different acquisition conditions and levels of control. As a result, each individual provided a total number of 240 fingerprint images to the database (10 prints × 12 samples/print × 2 sensors/sample). Figure 4 shows three examples acquired with the optical scanner under the 3 considered levels of control.

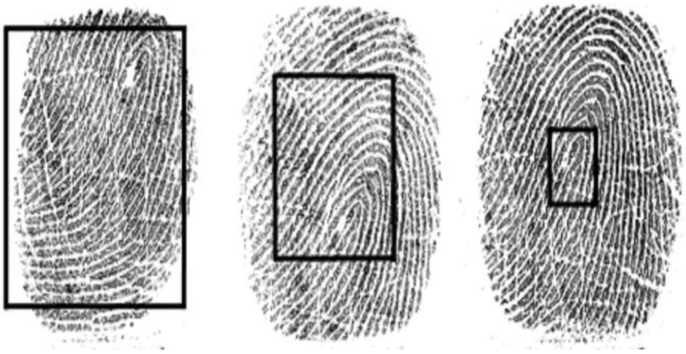

Fig. 4. Fingerprint images from MCYT corpus. Level of control from left to right: low, medium and high

Only the index fingers of the first 75 users in the database are used in the experiments. 10 print samples (optical scanner) per user are selected, 3 of them (each one from a different level of control) are used as expert training data and the other 7 are used as expert testing data. We have also considered a worst-case scenario using for each client the best 10 impostor fingerprint samples from a set of 750 different fingerprints.

All fingerprint images have been supervised and labelled according to the image quality by a human expert [29]. Basically, each different fingerprint image has been assigned a subjective quality measure from 0 (lowest quality) to 9 (highest quality) based on image factors like: incomplete fingerprint, smudge ridges or non uniform contrast, background noise, weak appearance of the ridge structure, significant breaks in the ridge structure, pores inside the ridges, etc. Figure 5 shows four example images and their labelled quality.

As a conclusion, each expert protocol comprises 75×7 client test attempts and 75×10 impostor test attempts in a near worst-case scenario.

8.2 Supervisor Protocol

Several methods have been described in the literature in order to maximize the use of the information in the training samples during a test [14]. For the error estimation in multimodal authentication systems, variants of the jackknife sampling using the leave-one-out principle are the common choice [23, 25]. In this work, and depending on the experiment at hand, one of the three following supervisor protocols has been used:

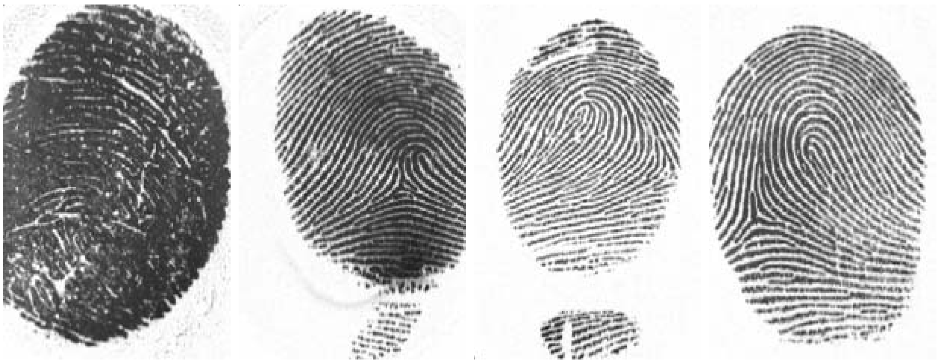

Fig. 5. Fingerprint images from MCYT corpus. Quality labelling from left to right: 0, 3, 6 and 9

Non-trained. All scores from client and impostor test attempts are used as supervisor test scores.

Trained-jacknife. One user is left out for supervisor testing, the supervisor training is carried out on the other users, the scheme is rotated for all the users and finally the errors are averaged.

Trained-bootstrap. N users are randomly selected with replacement for training, the testing is performed on the other users, the scheme is iterated B times and finally the errors are averaged.

8.3 Results

In the first experiment, we evaluate the verification performance of the three following fusion strategies: *i*) Sum Rule [12], which consists in averaging expert outputs; *ii*) The non-adaptive Bayesian Conciliation scheme [11] as described in section 5.1 (i.e. with $s_{ij} = 1$ for all authentication claims); and *iii*) The adaptive fusion strategy based on signal quality described in section 5.3. The non-trained supervisor protocol has been used for testing the Sum Rule approach whereas the trained-jacknife protocol has been followed for the other two trained fusion approaches. For the fingerprint expert, we have used the quality labels in MCYT database normalized into the range $[0, 2]$. For the speech expert $s_{ij} = 1$ is used. Trade-off verification results comparing the three fusion approaches are shown in Figure 6. As a result, any of the three fusion strategies clearly outperforms both the fingerprint (EER=4.55%) and the speaker expert (EER=4.32%). We also observe that the Sum Rule approach (EER=1.66%) is outperformed by the simplified Bayesian Conciliation scheme (EER=1.33%). The introduction of quality signals leads to further verification performance improvements in almost every working point (EER=0.94%).

In Figure 7, the client/impostor decision boundaries for one left-out user of the trained-jacknife supervisor protocol is depicted together with score maps

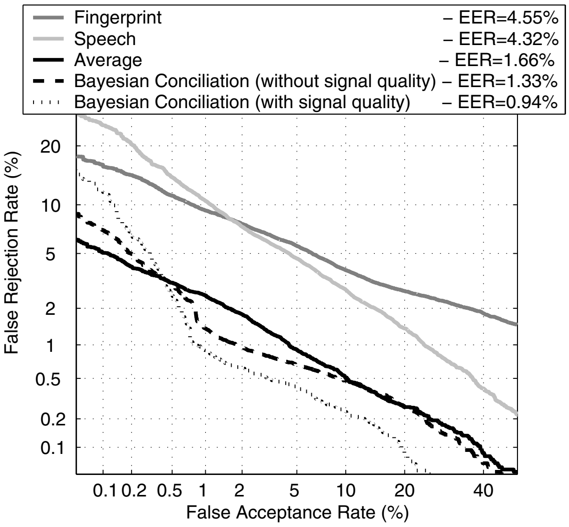

Fig. 6. Verification performance of fingerprint/speech experts and Sum/ Bayesian supervisors

of both training (background) and testing (enlarged) data. We note that the Sum Rule scheme does not take into account the actual client and impostor distributions, that is a skilled expert is weighted equally as a less skilled expert.

Some examples that may provide an intuitive idea about how the supervisor is adapted depending on the image quality of the input fingerprints are shown in Figure 8. We plot the decision boundaries for 2 different left-out users of the supervisor testing protocol together with score maps of both the training (background) and testing (enlarged) data. In the case the score quality is considered, we observe that the supervisor is adapted so as to increase or reduce the weight of the fingerprint expert opinion based on the fingerprint quality: the higher the image quality the higher the fingerprint expert weight and the lower the quality the lower the weight.

In the last experiment, we study the influence of an increasing number of clients N in the supervisor training set over the verification performance. In this case, the trained-bootstrap supervisor protocol with B=200 iterations has been used. As it is shown in Figure 9, the error rate decreases monotonically with the number of clients in the supervisor training set. In particular, a fast EER decay occurs for the first 10 clients and minor verification performance improvements are obtained for more than 20 users.

9 Conclusions

In this paper we have first summarized evidence that even one of the best known mono-modal recognition engines (human face recognition) is not able to reach a

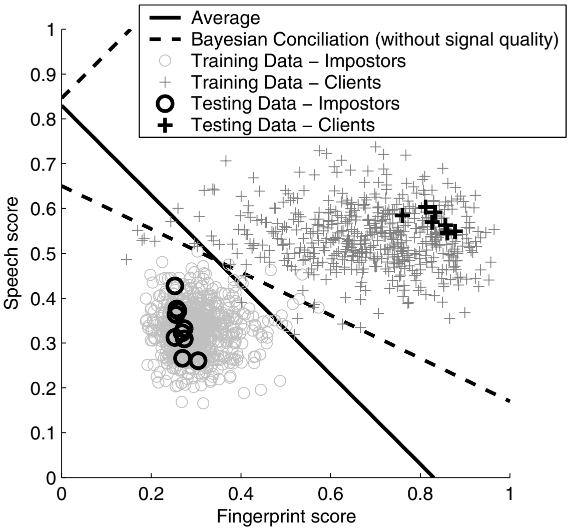

Fig. 7. Training/testing score maps and decision boundaries for Sum/Bayesian supervisors

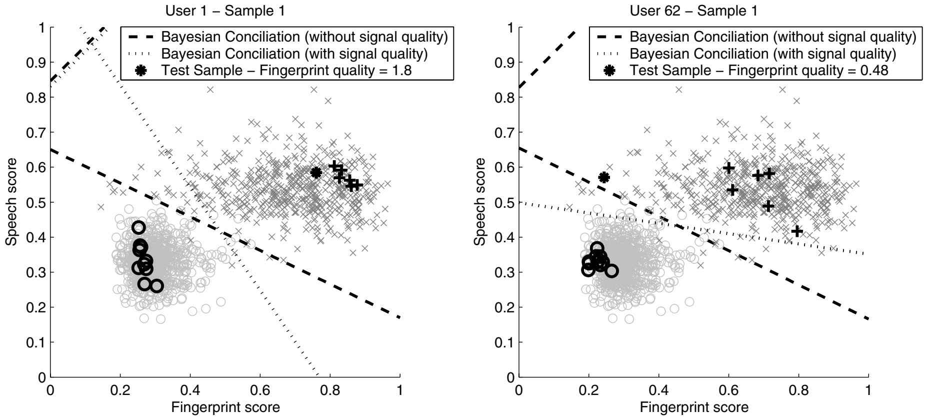

Fig. 8. Training/testing score maps and decision boundaries for Bayesian supervisors

recognition rate beyond 80 % when it is limited to a single view, i.e. a common approach in commercial applications. This has served as the motivation for, beginning with some common terminology and notations, the development of multi-modal automatic person authentication system models [11]. We have also explored an adaptive supervisor strategy and reviewed an implementation based

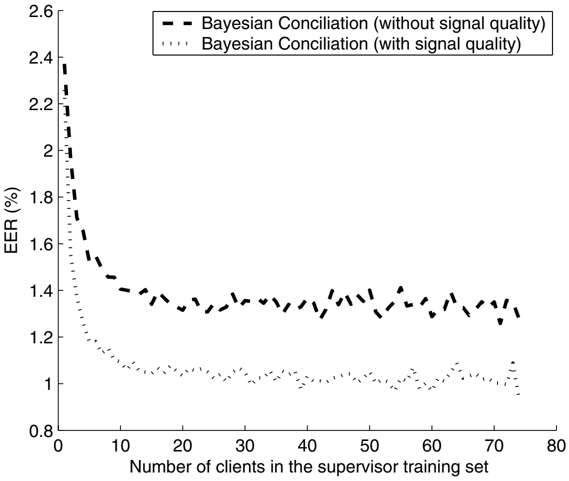

Fig. 9. Error rate vs number of clients in the supervisor training set

on signal quality of such a scheme [1]. The elements of a mobile authentication application based on speech and fingerprint data have been described and some experiments using this prototype on real data have been reported.

From the experiments, we conclude that multi-modal systems combining different biometric traits (EER=4.55% and EER=4.32% respectively for fingerprint and speaker experts in a near-worst case scenario) and using simple supervisor algorithms such as averaging can provide great benefits (EER=1.66%) in terms of verification error rates. Moreover, a Bayesian Conciliation fusion strategy have also been tested. In this case, it has been shown that weighting each expert output according to its past performance decreases error rates (EER=1.33%). Finally, we have also shown that the referenced adaptive fusion strategy can further improve the verification performance (EER=0.94%) compared to a trained but non-adaptive fusion strategy.

Future work includes the investigation of automatic quality measures for the different audio- and video-based biometric signals and the exploitation of the user-specific characteristics in the overall multi-modal authentication architecture.

10 Acknowledgements

In addition to Halmstad University, this work has been supported by the Spanish Ministry for Science and Technology under projects TIC2003-09068 and TIC2003-08382. J. F.-A. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research.

References

[1] Bigun, J., Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Multimodal biometric authentication using quality signals in mobile communications. In: Proc. of IAPR Intl. Conf. on Image Analysis and Processing, ICIAP, IEEE CS Press (2003) 2–13

[2] Baylis, G.C., Rolls, E., Leonard, C.M.: functional divisions of the temporal lobe neocortex. J. Neuroscience **7** (1987) 330–342

[3] Farah, M.J.: Is face recognition special? evidence from neuropsychology. Behavioral Brain Research **76** (1996) 181–189

[4] Elliott, E.S., Wills, E.J., Goldstein, A.G.: The effects of discrimination training on the recognition of white and oriental faces. Bulletin of the Psychonomic Society **2** (1973) 71–73

[5] Luce, T.S.: The role of experience in inter-racial recognition. Personality and Social Psychology Bulletin **1** (1974) 39–41

[6] Bothwell, R.K., Brigham, J.C., Malpass, R.S.: Cross-racial identification of faces. Personality and Social Psychology Bulletin **15** (1989) 19–25

[7] Bigun, J., Choy, K., Olsson, H.: Evidence on skill differences of women and men concerning face recognition. In Bigun, J., Smeraldi, F., eds.: Proc. of IAPR Intl. Conf. on Audio- and Video-based Person Authentication, AVBPA, Springer (2001) 44–51

[8] Kanade, T.: Picture processing system by computer complex and recognition of human faces. In: doctoral dissertation, Kyoto University. (1973)

[9] Atal, B.S.: Automatic recognition of speakers from their voices. Proceedings of the IEEE **64** (1976) 460–475

[10] Brunelli, R., Falavigna, D.: Person identification using multiple cues. IEEE Trans. Pattern Anal. and Machine Intell. **17** (1995) 955–966

[11] Bigun, E.S., Bigun, J., Duc, B., Fischer, S.: Expert conciliation for multi modal person authentication systems by bayesian statistics. In Bigun, J., Chollet, G., Borgefors, G., eds.: Proc. of IAPR Intl. Conf. on Audio- and Video-based Person Authentication, AVBPA, Springer (1997) 291–300

[12] Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. Pattern Anal. and Machine Intell. **20** (1998) 226–239

[13] Bigun, E.S.: Risk analysis of catastrophes using experts' judgments: An empirical study on risk analysis of major civil aircraft accidents in europe. European J. Operational Research **87** (1995) 599–612

[14] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern classification. Wiley (2001)

[15] Ben-Yacoub, S., Abdeljaoued, Y., Mayoraz, E.: Fusion of face and speech data for person identity verification. IEEE Trans. on Neural Networks **10** (1999) 1065–1074

[16] Verlinde, P., Chollet, G., Acheroy, M.: Multi-modal identity verification using expert fusion. Information Fusion **1** (2000) 17–33

[17] Gutschoven, B., Verlinde, P.: Multi-modal identity verification using support vector machines (SVM). In: Proc. of the Intl. Conf. on Information Fusion, FUSION, IEEE Press (2000) 3–8

[18] Ross, A., Jain, A.K., Qian, J.Z.: Information fusion in biometrics. In Bigun, J., Smeraldi, F., eds.: Proc. of IAPR Intl. Conf. on Audio- and Video-based Person Authentication, AVBPA, Springer (2001) 354–359

[19] Kittler, J., Messer, K.: Fusion of multiple experts in multimodal biometric personal identity verification systems. In: Proc. of the IEEE Workshop on Neural Networks for Signal Processing, NNSP. (2002) 3–12

[20] Duin, R.P.W.: The combining classifier: to train or not to train? In: Proc. of the IAPR Intl. Conf. on Pattern Recognition, ICPR, IEEE CS Press (2002) 765–770

[21] Roli, F., Fumera, G., Kittler, J.: Fixed and trained combiners for fusion of imbalanced pattern classifiers. In: Proc. of the Intl. Conf. on Information Fusion, FUSION. (2002) 278–284

[22] Jain, A.K., Ross, A.: Learning user-specific parameters in a multibiometric system. In: Proc. of the IEEE Intl. Conf. on Image Processing, ICIP. Volume 1. (2002) 57–60

[23] Fierrez-Aguilar, J., Ortega-Garcia, J., Garcia-Romero, D., Gonzalez-Rodriguez, J.: A comparative evaluation of fusion strategies for multimodal biometric verification. In: Proc. of IAPR Intl. Conf. on Audio- and Video-based Person Authentication, AVBPA, Springer (2003) 830–837

[24] Ellis, H.D., Ellis, D.M., Hosie, J.A.: Priming effects in childrens face recognition. British Journal of Psychology **84** (1993) 101–110

[25] Bigun, J., Duc, B., Fischer, S., Makarov, A., Smeraldi, F.: Multi modal person authentication. In et. al., H.W., ed.: Nato-Asi advanced study on face recogniton. Volume F-163., Springer (1997) 26–50

[26] Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ortega-Garcia, J.: Forensic identification reporting using automatic speaker recognition systems. In: Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP. Volume 2. (2003) 93–96

[27] Garcia-Romero, D., et al.: ATVS-UPM results and presentation at NIST'2002 speaker recognition evaluation (2002)

[28] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10** (2000) 19–41

[29] Simon-Zorita, D., Ortega-Garcia, J., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J.: Image quality and position variability assessment in minutiae-based fingerprint verification. IEE Proceedings Vision, Image and Signal Processing **150** (2003)

[30] Jain, A.K., Hong, L., Pankanti, S., Bolle, R.: An identity authentication system using fingerprints. Proceedings of the IEEE **85** (1997) 1365–1388

[31] Bigun, J., Granlund, G.H.: Optimal orientation detection of linear symmetry. In: First International Conference on Computer Vision, ICCV (London), Washington, DC., IEEE Computer Society Press (1987) 433–438

[32] Ortega-Garcia, J., et al.: MCYT baseline corpus: A bimodal biometric database. IEE Proceedings Vision, Image and Signal Processing **150** (2003)

[33] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M.: The DET curve in assessment of decision task performance. In: Proc. of ESCA Eur. Conf. on Speech Comm. and Tech., EuroSpeech. (1997) 1895–1898

[34] Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J.: Support vector machine fusion of idiolectal and acoustic speaker information in spanish conversational speech. In: Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP. Volume 2. (2003) 229–232

[35] Maio, D., Maltoni, D., Cappelli, R., Wayman, J.L., , Jain, A.K.: FVC2000: fingerprint verification competition. IEEE Trans. Pattern Anal. and Machine Intell. **24** (2002) 402–412