Multimodal Biometric Authentication using Quality Signals in Mobile Communications

Josef Bigun
Halmstad University, Sweden
Josef.Bigun@ide.hh.se

J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez
Universidad Politecnica de Madrid, Spain
{jfierrez, jortega, jgonzalez}@diac.upm.es

Abstract

The elements of multimodal authentication along with system models are presented. These include the machine experts as well as machine supervisors. In particular fingerprint and speech based systems will serve as illustration of a mobile authentication application. A novel signal adaptive supervisor, based on the input biometric signal quality is evaluated. Experimental results on data collected from mobile telephones are reported demonstrating the benefits of the proposed scheme[1].

1. Introduction

Automatic access of persons to services is becoming increasingly important in the information era. Although person authentication by machine has been a subject of study for more than thirty years [18, 1], it has not been until recently that the matter of combining a number of different traits for person verification has been considered [8, 5]. 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.). Within this technological framework, multimodal authentication is no longer conceived as a top security access control tool. 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 multibiometric system are described in [4] and [19], the former from a risk analysis perspective [3] and the later from a statistical pattern recognition point of view [9]. Both of them concluded (under some mild conditions which normally hold) 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 [2, 29, 15], 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 [26, 20] and a debate has come out investigating the benefits of both approaches [10, 25]. As an example, and within this debate, some researches have shown how to learn user-specific parameters in a trained fusion scheme [17, 11]. As a result, they have reported that the overall verification performance can be improved significantly.

In this work we focus on some other benefits of a trained fusion strategy. In particular, an adaptive trained fusion scheme is proposed and investigated 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 some definitions are provided. The elements of multimodal authentication along with major notations are introduced in section 3. In section 4, the statistical framework for con-

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

Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**

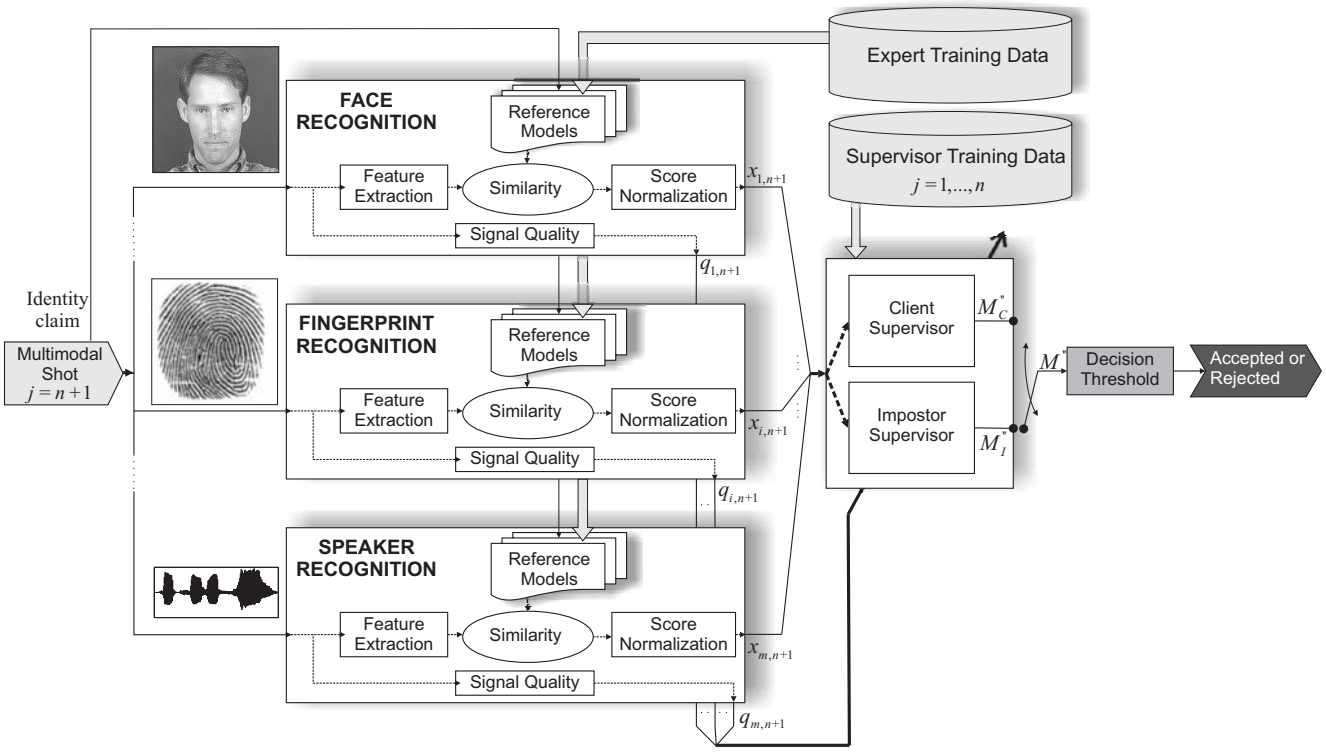

Figure 1. The proposed system model of multi-modal person authentication.

ciliating the different expert opinions together with the proposed adaptive fusion strategy is described. The components of our prototype mobile authentication application, namely fingerprint and speaker verification subsystems, are briefly described in section 5. Some experiments using the above-mentioned multimodal authentication prototype on real data are reported in section 7, where the benefits of the proposed adaptive fusion scheme are also explored. Conclusions will be finally given in section 8.

2. 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 pretended 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 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.

3. System model

Below is a list of the major notations we use throughout the paper, see also Figure 1.

i Index of the experts, $i \in 1 \dots m$

j Index of the shots, $j \in 1 \dots n, n+1$

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).

4. 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 [4] 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 1).

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 $M''_{\mathcal{C}} = E[Y_{n+1} | \mathcal{C}, x_{i,n+1}]$ (the prime notation will become apparent later on). In case of impostor supervisor, $M''_{\mathcal{I}} = E[Y_{n+1} | \mathcal{I}, x_{i,n+1}]$ 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 [4] is obtained, see [3] for further background and details. In the following, we summarize this algorithm in the two cases where it can be applied.

4.1. Simplified supervisor algorithm

When only the experts scores x_{ij} are 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}. First the client as well as impostor supervisors calibrate the experts according to their past performance, yielding (for the client supervisor)

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

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

The algorithm described above has been successfully applied in [6] 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.

4.2. 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

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

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

The algorithm described above has been successfully applied in [3] but not in a multimodal authentication application and combining only human expert opinions.

4.3. 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 4.2 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 can be used [14] 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 1). 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 the quality label of the biometric trait used by expert i in shot j and the average quality of the biometric samples used by expert i for modelling the claimed identity respectively. 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

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

5. Monomodal experts

5.1. Speaker expert

For the experiments reported in this paper, the GMM-based speaker expert from Universidad Politecnica de Madrid used in the 2002 NIST Speaker Recognition evaluation [12] has been used. Below we briefly describe the basics, for more details we refer to [24, 12].

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. Moreover, 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).

Pattern comparison. 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

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

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.

5.2. Fingerprint expert

For the experiments reported in this paper, the minutiae-based fingerprint expert described in [27] has been used. Below we describe the basics, for more details we refer to [16, 27].

Image enhancement. The fingerprint ridge structure is reconstructed according to: i) grayscale level normalization, ii) orientation field calculation, according to [7] 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 [23], the resulting binary image after image enhancement, the detected minutiae superimposed on the thinned image and the resulting minutiae pattern are shown in Figure 2.

Pattern comparison. 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.

Figure 2. Fingerprint feature extraction process

6. 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 [22], 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 [13], 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 described in [21] has been applied.

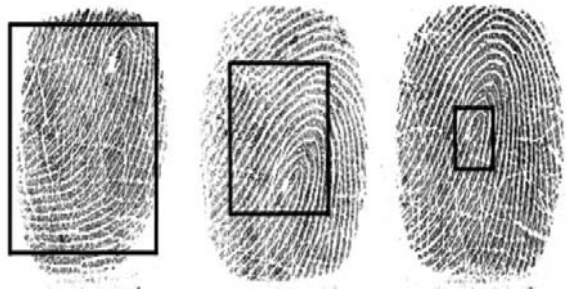

Figure 3. Fingerprint images from MCYT corpus. Control level from left to right: low, medium and high

7. Experiments

7.1. Database description and expert protocol

Cellular speech data consist of short utterances (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 (i.e. each impostor knew the mobile number and the way it was pronounced by the user he/she was forgering) per user are used as expert testing data. 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 [23]. Below, some information related to the experiments we have conducted is briefly described.

MCYT fingerprint subcorpus comprises 295 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 $\times$ 12 samples/print $\times$ 2 sensors/sample). Figure 3 shows three examples acquired with the optical scanner under the 3 considered control levels.

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 control level) 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 [28]. 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 4 shows four example images and their labelled quality.

Figure 4. Fingerprint images from MCYT corpus. Quality labelling from left to right: 0, 3, 6 and 9

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

7.2. Supervisor protocol

Several methods have been described in the literature in order to maximize the use of the information in the training

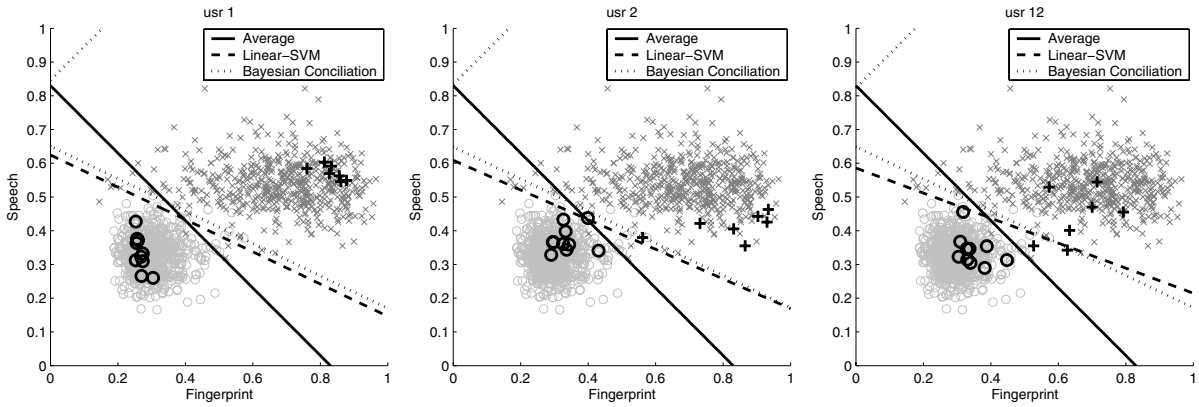

Figure 5. Training/testing scatter plot and separation surfaces for Sum/SVM/Bayes supervisors

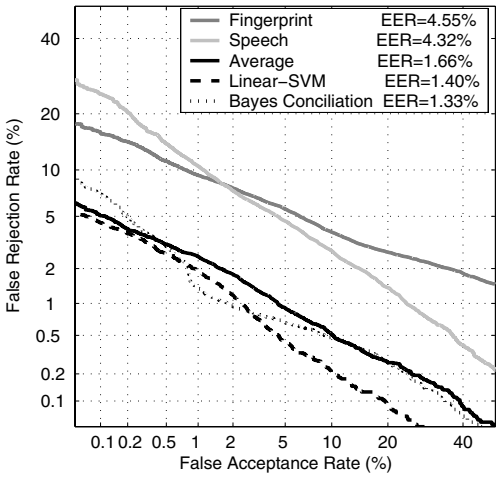

Figure 6. Verification performance of fingerprint/speaker experts and Sum/SVM/Bayes supervisors

samples during a test [9]. For the error estimation in multimodal authentication systems, variants of the jackknife sampling using the leave-one-out principle are the common choice [6, 11]. In this work, and depending on the experiment at hand, one of the three following supervisor protocols has been used:

Non-trained. In the case of a non-trained supervisor (e.g. Sum Rule), all expert test scores are used as supervisor test scores.

Trained-Jack. In the case we want to maximize the size of the training set, the jackknife paradigm is convenient. 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-Boot. In the case the size of the training set is varied, a bootstrap sampling is used: N users are chosen at random for training, the testing is performed on the other users, the scheme is iterated B times and finally the errors are averaged.

7.3. Results

In the first experiment, we evaluate the verification performance of the individual speaker and fingerprint experts and compare it to the verification performance of the 3 most popular multimodal fusion schemes, namely: i) Sum Rule [19], which consists of averaging expert outputs; ii) Support Vector Machine fusion [2] with linear kernel, which consists of separating client and impostor distributions by means of a large margin classifier (see [11] for details); and iii) The non-adaptive Bayesian Conciliation scheme [4] as described in section 4.1 (i.e. with $s_{ij} = 1$ for all authentication claims). The non-trained supervisor protocol has been used for testing the Sum Rule approach. The trained-Jack protocol has been followed for testing the other two fusion strategies. Verification results are plotted in Figure 6. We first observe that any of the three fusion strategies clearly outperforms both the fingerprint (EER=4.55%) and the speaker expert (EER=4.32%). We also note that both trained approaches outperform the Sum Rule (EER=1.66%). The SVM strategy (EER=1.40%) had the widest set of best performance working points but it was clearly inferior to the Bayesian scheme (EER=1.33%) at a large zone around the EER. This can be explained by equation (3) where impostor and client supervisor errors are equally important. Giving them non-equal weights (i.e. $w_{\mathcal{C}} \left|1 - M''_{\mathcal{C}}\right| - w_{\mathcal{I}} \left|0 - M''_{\mathcal{I}}\right| < 0$) is one way to shift the

Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)
0-7695-1948-2/03 $17.00 © 2003 **IEEE**

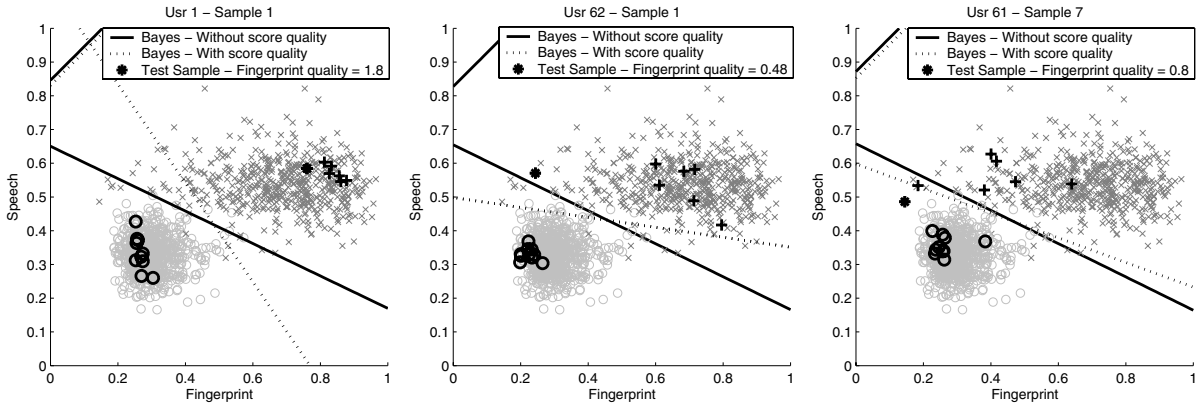

Figure 7. Training/testing scatter plot and separating surfaces for claim-adaptive Bayes supervisor

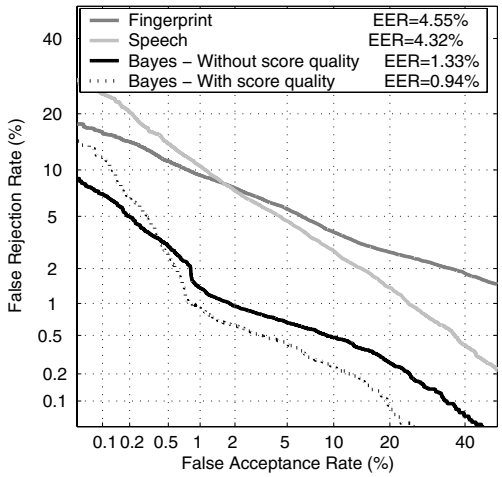

Figure 8. Verification performance of simplified and adaptive Bayes supervisors

optimal operation zone, where the Bayesian Conciliation outperforms the SVM approach. Nevertheless, we are only interested in setting a starting point for the following experiments, not to make a comparative study for which wider scope than that of this paper would be necessary.

In Figure 5, the client-impostor separation surfaces for 3 different left-out users of the trained-Jack supervisor protocol are depicted together with the score map of both the 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. Also interesting is the similarity between SVM and Bayesian Conciliation separation surfaces, the former being highly dependent on the data near the surface (as it is expected) while the later is almost the same for different users of the leave-one-out scheme. We stress the fact that different separation surfaces are obtained due to the jackknife procedure, although we are estimating the error for a trained but non-adaptive fusion scheme.

In the second experiment, we investigate the benefits of the proposed adaptive strategy described in section 4.3. 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. The trained-Jack supervisor protocol has been followed. Verification results comparing the non-adaptive scheme (without score quality) and the claim-adaptive case (with score quality) are shown in Figure 8. As a result, verification performance improves almost in every working point (EER decreases from 1.33% to 0.94%).

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 7. We plot the separation surfaces for 3 different left-out users of the supervisor testing protocol together with the score map 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 increasing the number of clients N in the supervisor training set over the verification performance. In this case, the trained-Boot 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 decay occurs for the first 10

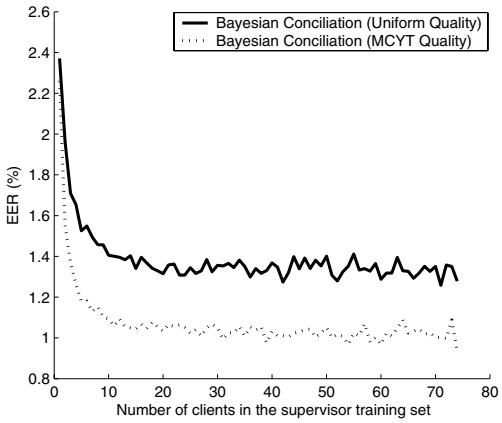

Figure 9. Error rate vs number of clients in supervisor training set

clients and small improvements are obtained for more than 20 users.

8. Conclusions

In this paper we have first reviewed some common terminology and notations in multimodal authentication systems and developed the statistical model proposed in [4]. We have also derived an adaptive supervisor strategy and proposed a signal quality-based implementation of such a scheme. 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 multimodal systems combining different biometric traits (EER=4.55% and EER=4.32% respectively for the 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, various referenced trained supervisors have been also 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 proposed 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 multimodal authentication architecture.

9. Acknowledgements

This work has been supported by the Spanish Ministry for Science and Technology under project TIC2000-1669-C04-01. J. F.-A. also thanks Consejeria de Educacion de la Comunidad de Madrid and Fondo Social Europeo for supporting his doctoral research.

References

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

[2] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz. Fusion of face and speech data for person identity verification. *IEEE Trans. Neural Networks*, 10(5):1065–1074, 1999.

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

[4] E. S. Bigun, J. Bigun, B. Duc, and S. Fischer. Expert conciliation for multi modal person authentication systems by bayesian statistics. In J. Bigun, G. Chollet, and G. Borgefors, editors, *Audio and Video based Person Authentication - AVBPA97*, pages 291–300. Springer, 1997.

[5] J. Bigun, G. Chollet, and G. B. Eds. *Proceedings of the first international conference on Audio and Video based Person Authentication - AVBPA97*, volume LNCS-1206. Springer, 1997.

[6] J. Bigun, B. Duc, S. Fischer, A. Makarov, and F. Smeraldi. Multi modal person authentication. In H. W. et. al., editor, *Nato-Asi advanced study on face recogniton*, volume F-163, pages 26–50. Springer, 1997.

[7] J. Bigun and G. H. Granlund. Optimal orientation detection of linear symmetry. In *First International Conference on Computer Vision, ICCV (London)*, pages 433–438, Washington, DC., June 8–11 1987. IEEE Computer Society Press.

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

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001.

[10] R. P. W. Duin. The combining classifier: to train or not to train? In *Proc. of the 16th Intl. Conf. on Pattern Recognition - ICPR 2002*, pages 765–770, 2002.

[11] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero, and J. Gonzalez-Rodriguez. A comparative evaluation of fusion strategies for multimodal biometric verification. In *Audio and Video based Person Authentication - AVBPA 2003*, pages 830–837. Springer, 2003.

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

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

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

[15] B. Gutschoven and P. Verlinde. Multi-modal identity verification using support vector machines (SVM). In *Proc. of the 3rd Intl. Conf. on Information Fusion*, pages 3–8, 2000.

[16] A. K. Jain, L. Hong, S. Pankanti, and R. Bolle. An identity authentication system using fingerprints. *Proceedings of the IEEE*, 85(9):1365–1388, 1997.

[17] A. K. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *Proc. of the 2002 IEEE Intl. Conf. on Image Processing - ICIP 2002*, volume 1, pages 57–60, 2002.

[18] T. Kanade. Picture processing system by computer complex and recognition of human faces. In *doctoral dissertation, Kyoto University*. November 1973.

[19] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. and Machine Intell.*, 20(3):226–239, 1998.

[20] J. Kittler and K. Messer. Fusion of multiple experts in multimodal biometric personal identity verification systems. In *Proc. of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 3–12, 2002.

[21] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, , and A. K. Jain. FVC2000: fingerprint verification competition. *IEEE Trans. Pattern Anal. and Machine Intell.*, 24(3):402–412, 2002.

[22] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of decision task performance. In *Proc. of ESCA 5th Eur. Conf. on Speech Comm. and Tech. - EuroSpeech '97*, pages 1895–1898, 1997.

[23] J. Ortega-Garcia et al. MCYT: A multimodal biometric database. In *Proc. of the COST-275 Workshop: The Advent of Biometrics on the Internet*, pages 123–126, 2002.

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

[25] F. Roli, G. Fumera, and J. Kittler. Fixed and trained combiners for fusion of imbalanced pattern classifiers. In *Proc. of the 5th Intl. Conf. on Information Fusion - FUSION 2002*, pages 278–284, 2002.

[26] A. Ross, A. K. Jain, and J. Z. Qian. Information fusion in biometrics. In J. Bigun and F. Smeraldi, editors, *Audio and Video based Person Authentication - AVBPA 2001*, pages 354–359. Springer, 2001.

[27] D. Simon-Zorita, J. Ortega-Garcia, M. Sanchez-Asenjo, and J. Gonzalez-Rodriguez. Facing position variability in minutiae-based fingerprint verification through multiple references and score normalization techniques. In *Audio and Video based Person Authentication - AVBPA 2003*, pages 214–223. Springer, 2003.

[28] D. Simon-Zorita, J. Ortega-Garcia, M. Sanchez-Asenjo, and J. Gonzalez-Rodriguez. Minutiae-based enhanced fingerprint verification assessment relaying on image quality factors. In *Proc. of the 2003 IEEE Intl. Conf. on Image Processing - ICIP 2003*, 2003.

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