(C) In Proc. Face Recognition: From Theory to Applications*, Nato Advanced Study Institute (ASI) Programme, Stirling, UK, June 23 - July 4, 1997, Springer, Lecture Notes in Computer Science series*

Multi-Modal Person Authentication

J.Bigün[1], B. Duc[2], F. Smeraldi[1], S. Fischer[2], and A. Makarov[1]

[1] EPFL Microprocessor and Interface Laboratory CH-1015 Lausanne
[2] EPFL Signal Processing Laboratory CH-1015 Lausanne

Abstract. This paper deals with the elements of a multi-modal person authentication systems. Test procedures for evaluating machine experts as well as machine supervisors based on leave-one-out principle are described. Two independent machine experts on person authentication are presented along with their individual performances. These experts consisted of a face (Gabor features) and a speaker (LPC features) authentication algorithm trained on the M2VTS multi-media database. The expert opinions are combined yielding far better performances by using a trained supervisor based on Bayesian statistics than individual modalities aggregated by averaging.

1 Introduction

Person authentication has been gathering considerable interest due to the easy access to computers and communication technologies. Recently, the audio and video based authentication techniques have been jointly used [5] in an attempt to find dependable solutions for the challenging problem of person authentication. The need for multi-modality is motivated by the fact that the speech and image based mono-modal authentication technologies are starting to reach a performance saturation.

With the increase of computation performance, authentication using multi-modalities, in particular vision and sound, is becoming more realistic. A fundamental reason for multi-modality is the inherent limitations of the information in a single modality. Biological systems tend to solve the problem by using multiple cues. It is more difficult to find people who resemble each other pictorially *and* vocally than for example to find people who resemble each other only pictorially. Consequently, the multi-modal authentication is helped by this low prior-probability. We investigate two modalities to be used in person authentication.

However, using multi-modal techniques require an automatic mechanism, a *machine supervisor*, for conciliating (sometimes contradictory) machine "opinions" to a single and more reliable opinion. It also requires test procedures for evaluation of algorithms constituting the machine experts and the machine supervisor which delivers a joint opinion by calibrating and aggregating the expert opinions, [1]. The supervisor algorithm used here is based on [2] which was originally developed for human experts assessing the risks for rare events such as catastrophes. This is motivated in that erroneously rejecting a client of a system or accepting an impostor can be assumed to be a rare event for a machine expert, as they are designed to reduce the risk of these events.

Here we describe the elements of multi-modal person authentication by using speech and face sensors which are not perceived as intrusive by their users. The works of [8,12,19] share conceptually similar interests with this paper.

2 System Model and Definitions

2.1 Identification versus Authentication

Person authentication and *person identification* are of primary interest for a number of security applications. Both will be briefly summarised as there is an important distinction, which has practical consequences, between the two.

In *authentication* applications, the clients are known to the system whereas the impostors can potentially be the world population. In such applications the scenario is cooperative, that is the users provide their pretended identities which are known to the system. In case the candidate provides an unknown identity, he will be rejected without further check. Authentication is the focus of the concepts developed in this work, although many of these are also useful for person identification.

In *identification* applications, the scenario is non-cooperative and therefore there is no identity claim. The situation is very much like that of a database query. The candidate is compared to the entire database, and the correct identity should be among the best matches. This is the simplest form of identification which is also called *closed-universe identification.* In the more elaborate versions of the identification, the candidate may or may not belong to the database. In the case of the latter, the system should detect this and reject the query in order to reduce the identification error. This is called *open-universe identification.* The rejection process in open-universe identification systems is an implicit authentication step.

2.2 Supervisor and Experts

We have a system consisting of one supervisor and m experts. The supervisor does not interfere with the computational processes of the experts. It only asks the experts their opinions about the claims of a candidate. Below is a list of the major notations we use throughout the paper, see also Figure 1. Other notations only important for a module are described in place.

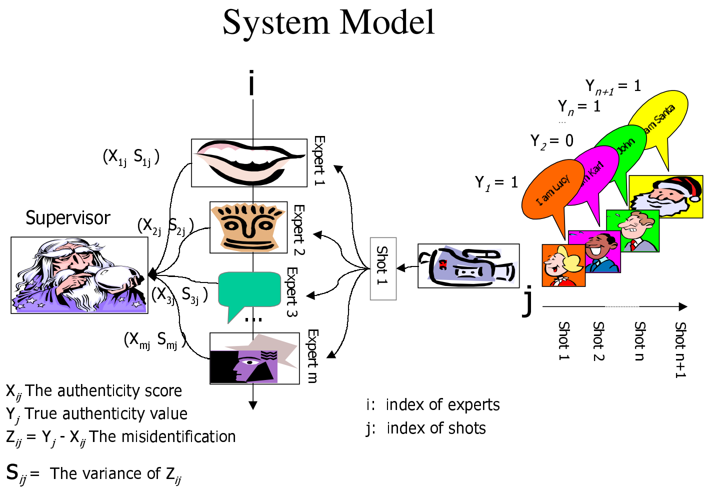

Fig. 1. The system model of multi-modal person authentication.

A take: A data package (e.g. speech+image) of a candidate, without identity claim

A shot: A data package of a candidate with an identity claim

i, j: Indices of the experts, $i \in 1 \cdots m$, and of the shots, $j \in 1 \cdots n, n+1$.

K, L: The number of persons and takes in a database

X_{ij}: Authenticity score delivered by expert i on shot j's being a take of the claimed client

s_{ij} The variance of X_{ij} as estimated by expert i. The experts are allowed to provide a quality of the score which is modelled to be inversely proportional to s_{ij}.

Y_j The true authenticity score of shot j's being that of a client.

Z_{ij} The error score of an expert $Z_{ij} = Y_j - X_{ij}$

T: Hard decision threshold (accept or reject)

Single variable indices, e.g. s_j, P_j, Z_j represent aggregated (supervisor) variables instead of expert variables. In the context of supervisor design, we assume that the shots $1 \cdots n$ are new shots of the clients, i.e. the experts have trained on other shots of the corresponding clients. Shot $n+1$ is the shot of a candidate which neither the experts nor the supervisor have trained on. Therefore shot $n+1$ can be considered to belong to a future instant, or an instant when the system is in full use. During supervisor training, we also assume that the training phase of the experts is already achieved.

3 Evaluation

3.1 Methodology

The machine opinions are pairs of (X_{ij}, s_{ij}). They are originally in form of distances to the reference model. But for comparison purposes these are mapped to the $]0,1]$ interval by the experts themselves. The a priori threshold for separating acceptance and rejection is assumed to be 0.5. By varying it in the $]0,1]$ interval, one can influence two types of error rates: for example, if the threshold increases, the *false acceptance* (FA) rate decreases, but the *false rejection* (FR) rate increases. Several ways of displaying the behaviour of the error rates are possible. Receiver Operating Characteristics (ROC) curves show the false acceptance versus the false rejection. The threshold value is an implicit parameter of the curve. The terminology is taken from radar technology where the problem is to detect a target. When comparing two such curves, the one closest to the axes corresponds to the best method. However, the sensitivity of a point on the ROC curve with respect to the threshold is not possible to view as the threshold is implicit. This is sometimes desirable since threshold explicit curves reveal how easy it is to find the best operational threshold. By varying the threshold, one can reach a point where the FA and FR rates take the same value. This value, called Equal Error Rate (EER), provides a way to characterise a method with a single number, allowing a quick comparison.

Normally, by using a threshold T, the inequality $X_{i,j} > T$ can be turned to a decision of *accept* when fulfilled, or to a decision of *reject* otherwise. We rewrite this inequality by subtracting it from y_j yielding $Z_{i,j} < y_j - T$. The inequality yields an acceptance decision when it is fulfilled, rejection otherwise. Therefore an acceptance is a false acceptance when the inequality is fulfilled for $y_j = T_f = 0$. Likewise a rejection decision represents a false rejection when the inequality is *not* fulfilled for $y_j = T_t = 1$:

$$\text{False Acceptance} \Leftrightarrow Z_{i,j} < T_f - T = -T \tag{1}$$

$$\text{False Rejection} \Leftrightarrow Z_{i,j} \geq T_t - T = 1 - T \tag{2}$$

Consequently, the integral of the frequency function of Z taken over the semi-axes defined by (1) and (2) represent the FA and FR functions. To be more precise

$$FA(T) = \int_{z<T_f-T} f(z)/C dz = F(T_f - T) \quad (3)$$

$$FR(T) = \int_{z\geq T_t-T} f(z)/C dz = 1 - F(T_t - T) \quad (4)$$

where $f(z)$ is the frequency of Z, and C is a normalisation constant so that F, the integral of f, is a distribution function i.e. $F(\infty) = 1$.

This conclusion is interesting since f can be estimated via the histogram of Z in practice. As both histogram and summation (integral) routines are widely available in computer environments, the implementations of $FA(T)$ and $FR(T)$ computations are particularly simple, as compared to a straightforward approach, in which T's must be varied. This approach is also the one adopted here. FA, FR and the Total Error TE, which is FA + FR, are functions of the same threshold. The FA and FR discussion is valid for both experts and supervisors.

3.2 Test Protocols for Experts and Supervisors

Authentication algorithms need to be compared. For this reason databases which represent realistic situations should play a central role in evaluating verification technologies. The M2VTS database, [21], is a digital multi-media person database which, to the extent limited by storage requirements, takes into account the demands of current speech and image based authentication technologies. This database contains speech and video data of speaking persons and images representing head rotations of each person. Due to storage requirements, the speech is restricted to utterances of the digits 0..9. The database is made up from $K = 37$ different people and provides $L = 4$ takes for each person. The takes were recorded at one week intervals or when drastic face changes occurred in the meantime. During each take, people have been asked to count from '0' to '9' in their native languages (most of the people are French speaking).

For evaluation experiments, a database should ideally be split into three subsets: a *training set*, which is used for designing the system, an *evaluation set*, which is used for determining thresholds and which should consist of data independent from the training set, and a *test set* for estimating the performance of the system, i.e. the system is completely determined and works in the authentication mode. The error rates are estimated on the test set. Here, as few persons are available due to the nature of the multi-modal authentication, we prefer not to consider an evaluation set, and use a priori chosen thresholds for a functioning system.

Several methods have been described in the literature in order to maximise the use of the information in a database during a test [16]. However, it appears that only variants of bootstrap sampling are relevant for applications such as authentication. The details of our *Expert Protocol*, which has similarities with the Jack knife sampling and uses the leave-one-out principle, is given below.

The experiments have been conducted by leaving out both one person (all her takes) and one take (all persons). Alternatively, each person is labelled as an impostor, while the $K-1$ (i.e. 36) others are considered as *clients*. For each combination, $L-1$ (i.e 3) takes of the $K-1$ clients build the training set and the L'th take (i.e. 4'th) series is used as evaluation set in the following way: each client tries to access under the correct identity, and the impostor tries to access under the identities of the $K-1$ clients. This makes $K-1$ authentic tests and $K-1$ impostor tests. The procedure is repeated L times, by considering each take as the test series alternatively. In total, the client and impostor verification amount each to $K \times L \times (K-1)$, which evaluates to 5328 shots for the M2VTS database. Testing impostor access with persons belonging to the training set has not been used, as it is considered too easy to discriminate between persons present in the training set, even if the data themselves are not present in the training set. The Expert Protocol is summarised in Figure 2. All experts are supposed to give their opinions on a particular shot, according to the Expert Protocol. Therefore for a given expert such an opinion description can be unambiguously represented by a tuple

$$(\mathrm{ET\_LABEL}, \mathrm{C\_ID}, Y_j, X_{i,j}, P_{i,j}) \tag{5}$$

where ET_LABEL and C_ID represent the unique identities of the expert training set and the claimed identity (of a client). For simplicity we use the combination of the left-out person identity and the left out take in order to obtain the ET_LABEL. For example the tuple $(\mathrm{BP\_04}, \mathrm{CC}, 0, 0.4, 0.8)$ represents an impostor (since $Y_j = 0$, this is an impostor claim and ET_LABEL=BP_04 reveals that the actual identity of the person is BP) trial which obtained the score of 0.4, and the quality of the score 0.8.

Such opinions and ground truths are used to estimate the two main performance characteristics of the authentication, namely the FA rate and the FR rate of an expert. However, assuming that the experts deliver their opinions according to the Expert Protocol, we need another test procedure for evaluating the performance of the multi-modal system. We call this protocol, the *Supervisor Protocol* as it is slightly different than the Expert Protocol, even though it embraces the same principle.

The Supervisor Protocol uses the opinions and the ground truth delivered by the Expert Protocol for each expert, as given by (5), yielding 2×5328 such tuples for the M2VTS database. This leaves out all opinions related to the identity of a single person. That is, tuples with ET_LABEL or C_ID (whichever contains the identity of the person to be left out) are left out. The left-out opinions are used to test the supervisor, while the inliers are used to train the supervisor. Consequently the supervisor training set, ST_LABEL, can be represented by using the left out identity.

To fix the ideas let the left out person be ST_LABEL=BP. As a result of the procedure, there are $4 \times L \times (K-1) = 576$ opinions in the test set which leaves $2 \times K \times L \times (K-1) - 4 \times L \times (K-1) = 10080$ opinions for the training set by leaving out all BP related opinions from the Expert Protocols. By rotating the left out person one can obtain $K = 37$ training and test sets yielding $4 \times K \times L \times (K-1)$ expert opinion descriptions represented by tuples:

$$(\text{ST\_LABEL}, \text{C\_ID}, Y_j, X_j, P_j) \tag{6}$$

in which the supervisor and the expert training sets have no, or reasonable dependencies. These can be aggregated by taking their means in order to compute FA and FR curves for a supervisor. A complementary way of testing two supervisor performances independent of the experts relies on simulating expert opinions, see [3].

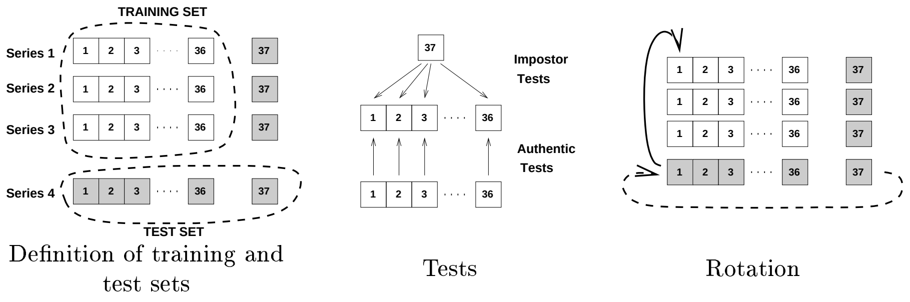

Fig. 2. Expert Protocol. The database is divided into a training set of $L-1=3$ takes of $K-1=36$ persons and a test set consisting of all K persons. Each configuration brings a total of $K-1$ authentic accesses, and $K-1$ impostor accesses.

4 Face Expert

Attributed graphs describe objects on sparse locations, by attaching to each node a feature vector that contains information on the local neighbourhood of the node location. Here, we use the modulus of complex Gabor responses as features, from filters with 6 orientations and 3 resolutions. For a discussion on their usefulness for image analysis applications, see [6,10].

4.1 Elastic Graph Matching

Each face is represented by a set of feature vectors positioned on nodes of a coarse, rectangular grid placed on the image. Comparing two face images is accomplished by matching and adapting a grid taken from one image to the features of the other image [20].

Elastic Graph Matching (EGM) consists in locating an attributed graph on the image that is as close as possible to the reference graph. The distance between two graphs is evaluated by a distance function, that considers both the feature vectors of each node and the deformation information attached to the edges.

We consider distance measures where the contribution from nodes and edges are independent, more precisely:

$$d(\Gamma, R) = \sum_{i=1}^{N_n} d_n(\Gamma_{n_i}, R_{n_i}) + \lambda \sum_{j=1}^{N_e} d_e(\Gamma_{e_j}, R_{e_j}), \tag{7}$$

where Γ_{n_i} and R_{n_i} represent the Gabor feature vectors at the ith node of the test and reference grids respectively. The edge vector of the test and reference grid are represented by Γ_{e_j} and R_{e_j}. N_n, N_e are the number of nodes and edges, respectively. λ is a weighting factor which characterises the stiffness of the graph. A *plastic* graph which opposes no reaction to deformation corresponds to $\lambda = 0$, while a totally rigid graph corresponds to the limit case $\lambda = \infty$.

The matching procedure consists of two consecutive steps [20]. The first step is used for obtaining an approximate match, with a rigid grid, which is equivalent to setting a high value of λ. Starting from this initial guess, the grid is deformed in order to minimise (7).

4.2 Coarse-to-Fine Matching

The computation of a feature vector for a node at a given location requires a filtering operation for each feature. If Gabor filters are used with 6 orientations and 3 resolutions, 18 filtering operations are required. As this can be computationally demanding, we suggest the use of coarse to fine matching when doing the rigid Gabor response matching.

For a graph matching, the filter responses may be needed only on a reduced subset of points in the image. Depending on the number of points visited, it may become computationally less expensive to compute the Gabor responses only at required points by convolution in the spatial domain.

We consider a multi-resolution description of the image. First, the lowest resolution image is considered for matching. As a consequence, the objective function is smoothed, and the matching may be undertaken on a sub-sampled lattice. This property is intimately related to the fundamental sampling theorem: as the objective function has been low-pass filtered, it may be sampled at a coarser step without loss of information. Figure 3 shows that the low-resolution image provides a smooth objective function, while the high frequency information generates a forest of local minima. However, the minima are more precisely localised when the high frequencies are incorporated. Consecutive refinements are obtained by incorporating higher resolution information and by searching on a finer grid around the current estimate.

Coarse-to-fine strategies may get trapped in local minima [15]. A remedy to this weakness consists in the elaboration of mixed fine-to-coarse and coarse-to-fine strategies. However, if only one head is present and occupies a significant part of the image, we noticed that this problem does not occur.

In practice, a Gaussian pyramid is built [9]. In a pyramidal implementation, the size of images depends on the resolution. The pyramid is built recursively, by

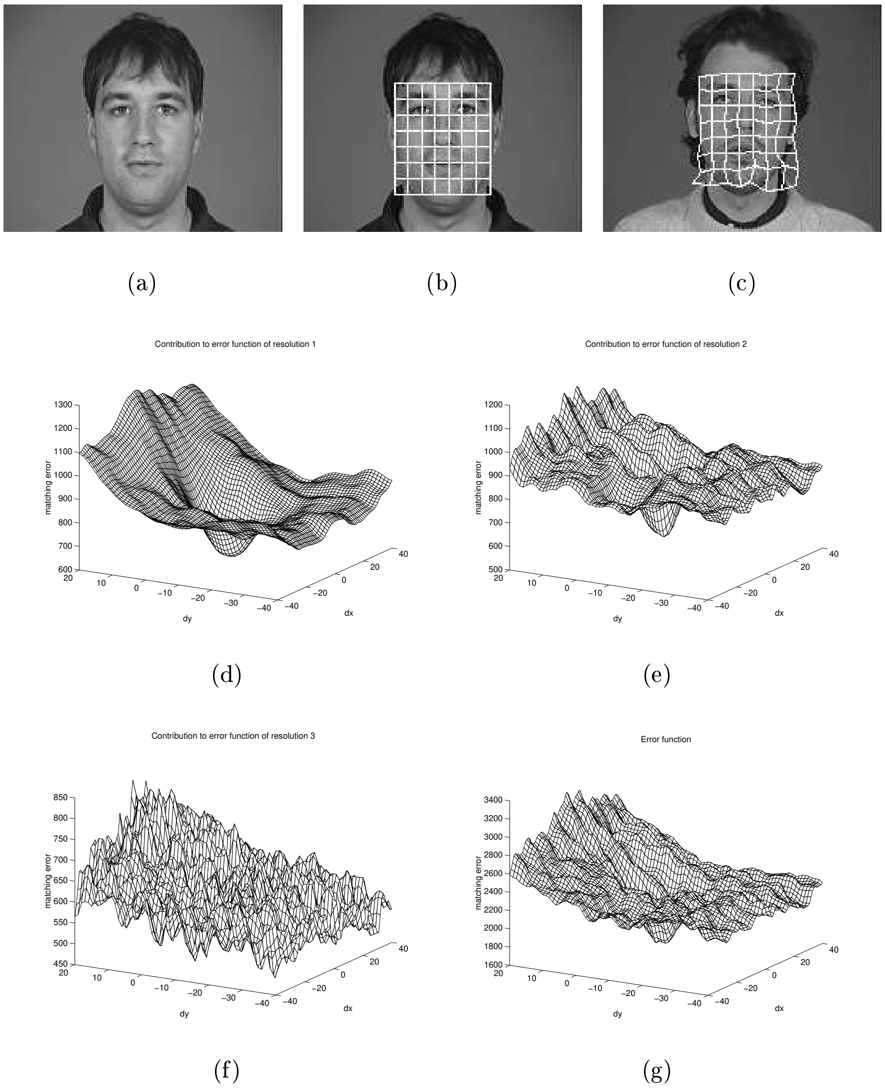

Fig. 3. Objective function for the rigid translation of the graph on the search window. Gabor responses on three resolutions are used. Here, the contribution of each resolution is shown separately. (a) original image from which the reference graph is taken. (b) reference graph superimposed on the reference image. (c) matched graph superimposed on the test image. (d) objective function with only the lowest resolution. (e) objective function with only the medium resolution. (f) objective function with only the highest resolution. (g) total objective function. While low resolution provide smooth, convex objective functions, high resolution responses provide sharper minima, and are used for refinement.

building a new, lower-resolution level from the previous one by low-pass filtering and sub-sampling with a factor two, so that the size of the image is divided by two at each iteration. The low-pass filtering was achieved with separable Gaussian filters. With the pyramidal implementation, the lattice spacing in *pixels* is kept constant through all levels: a displacement of 1 pixel at level n of the pyramid corresponds to a displacement of 2^n pixels in the level 0 of the pyramid, which is the original image.

The definition of filters is simplified by defining a set of filters for a single resolution and a complete set of orientations. These filters are applied to each level of the pyramid, to obtain a complete set of resolutions. A significant reduction of the amount of computations is obtained for low frequency responses, compared to filtering the original image, as bandpass filters selecting low frequencies have a large support.

4.3 Dimensionality Reduction and Local Discriminants

The first step of the authentication consists in matching the image with the prototype grid of the claimed class (in the following, each person in the database is considered as a *class* of the classification problem). This prototype is taken as the mean of the feature vectors provided by all images of the considered person in the training set. It is expected that if the claimed identity is correct, the feature vector will be close to the prototype of the class; in case of an impostor, the matching will perform poorly. Unfortunately, early experiments showed that the *Residual Matching Error* (RME), i.e. $d(G, R)$ after matching with $\lambda = 0$, is not sufficient to discriminate between an impostor and the authentic person, see Section 4.4. This is partly due to the presence of noise in the measurement, but also due to the fact that not all nodes are discriminative. Indeed, the feature space considered here is very large: for an 8 by 8 grid comprising 18 Gabor responses at each node, a total of $N_G = 1152$ features is obtained.

Reducing the dimensionality is an efficient way to reduce the influence of noise [11,4]. From a training set consisting of several frontal views of each person, one establishes subspaces which maximise the dispersion of all classes while minimising the dispersion within the classes.

However, the number of training samples is small compared to the number of features. Also, the features on two graph nodes may be considered as independent. Therefore, it is reasonable to address dimensionality reduction independently at each node of the graph. If features are considered locally, the number of training samples is larger than the dimension of the feature space, which allows to apply feature reduction methods.

Local Discriminants Suppose that the dimensionality of the considered feature space is small compared to the number of training elements in each of the c considered classes. One would like to establish a decision criterion for the acceptance or rejection of the candidate. This criterion should be "small" if the candidate is the right person, and "large" in case of an impostor. Obviously, this decision has to be made on the difference between the prototype of the claimed class and the measured feature vector. The components of this difference do not bear the same significance, as some may be more relevant than others for the given class. Therefore, we propose the following discriminant criterion:

$$d_k(\boldsymbol{r}) = \left(\sum_{i=1}^{N_g} v_{k_i} (r_i - \mu_{k_i}) \right)^2 = \left(\boldsymbol{v}_k^t (\boldsymbol{r} - \boldsymbol{\mu}_k) \right)^2 \tag{8}$$

for class $k, k = 1...c$, where r_i are the components of the measurement vector $\boldsymbol{r}$, N_g is the dimension of the local feature space. Here, the local spaces are chosen as the sets of all orientations for a given resolution at a given node, so that $N_g = 6$. $\boldsymbol{\mu}_k$ is a mean of vectors $\boldsymbol{r}$ averaged over a set that will be precised in (9). The unknown coefficient vector $\boldsymbol{v}_k$'s are determined on the training set by minimising the ratio:

$$\begin{aligned} D_k &= \frac{\sum_{\boldsymbol{r} \in S_k} d_k(\boldsymbol{r})}{\sum_{\boldsymbol{r} \in (S - S_k)} d_k(\boldsymbol{r})} \\ &= \frac{\sum_{\boldsymbol{r} \in S_k} \boldsymbol{v}_k^t (\boldsymbol{r} - \boldsymbol{\mu}_k)(\boldsymbol{r} - \boldsymbol{\mu}_k)^t \boldsymbol{v}_k}{\sum_{\boldsymbol{r} \in (S - S_k)} \boldsymbol{v}_k^t (\boldsymbol{r} - \boldsymbol{\mu}_k)(\boldsymbol{r} - \boldsymbol{\mu}_k)^t \boldsymbol{v}_k} \\ &= \frac{\boldsymbol{v}_k^t W \boldsymbol{v}_k}{\boldsymbol{v}_k^t B \boldsymbol{v}_k}, \end{aligned} \tag{9}$$

where S_k is the set of training vectors belonging to class k, S is the whole training set, so that $(S - S_k)$ is the set of all impostors for class k. Here, $\boldsymbol{\mu}_k$ is the mean on S_k. By this, we are back to a two-class classification problem, where the classes are S_k and $(S - S_k)$. This formulation leads to a generalised eigenvalue problem: $W\boldsymbol{v}_k = \lambda B\boldsymbol{v}_k$, and $\boldsymbol{v}_k$ is given by the eigenvector corresponding to the smallest generalised eigenvalue. This is very similar to Fisher's discriminant ratio [11].

All local responses have to be combined in order to provide a unique, global dissimilarity measure for the considered face. Here, we build the global response by simply adding the contributions from the local discriminants. This discriminant measure will be abbreviated as LD.

Separation Parameters It is necessary to choose a threshold for defining acceptance/rejection intervals in the domain of possible responses from training data. Here we assume that the system will provide a soft decision between $[0, \infty[$, therefore a mapping between the original response interval and the interval $]0, 1]$ is needed.

A natural invertible mapping from $[0, \infty[$ to $]0, 1]$ is provided by the hyperbolic tangent function. For our purpose, the soft score $S \in]0, 1]$ should be 1 for an identity claim acceptation, and 0 for an identity claim rejection, whereas the global discriminant value tends to 0 for a perfect matching and to infinity for a maximum mismatch. We suggest the mapping:

$$S(x) = \tanh\left(\frac{\log(3)}{2x}t\right) \tag{10}$$

where t is an empirically chosen constant. Since by definition $S(t) = 0.5$, t will be called the *Separation Parameter* (SP), as it acts like a decision point on x between acceptance and rejection intervals. In the case of a soft decision the SP acts as a parameter selecting the mapping function. In the case of a hard decision SP is simply the threshold. We have chosen it as the minimal distance measure among the training impostors.

4.4 Experiments with Face Authentication

Local Feature Reduction for Authentication In order to motivate the process of dimensionality reduction, we first want to show that the Euclidean distance between features, i.e. the residual matching error $d(G, R)$ with $\lambda = 0$, is not sufficient for a reliable decision. Figure 4 shows as an example distances of training and test samples with person 15 used as reference. It turns out that the distance to the reference view is clearly not sufficient to detect impostors.

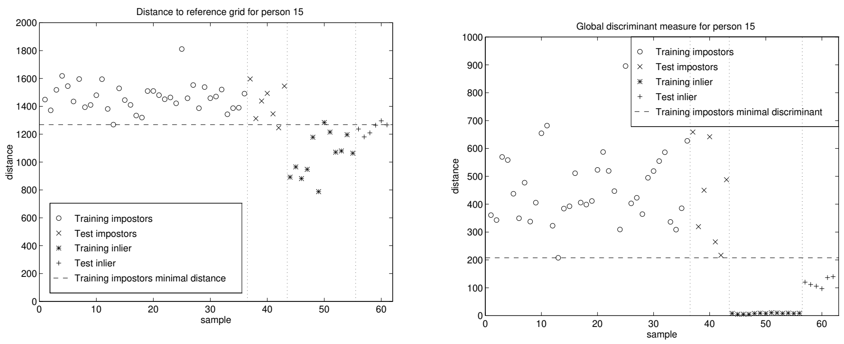

Fig. 4. Plot of distances for person (or class) 15. The distance of the grids of different kind of images, namely impostors in the training and the test set, members of the class in the training and test set, are shown. If one uses the minimal distance on the training impostors as a threshold for the decision, some members of the class in the training and test set are misclassified if the residual matching error is used (Left), whereas all members of the class are correctly classified with the local discriminants (Right).

A representation of discriminant values for the same person is also shown in the same figure. Now the discrimination of impostors is much more powerful. One can notice that there seems to be some over-training, as the discrimination measure is almost zero for all members of the considered class in the training set, and significantly larger for images of the same class in the test set, while remaining smaller than the threshold. This is due to the small number of training samples for each person in the database.

At that point, the discriminant values in the $[0, \infty[$ interval are normalised to the $[0, 1]$ interval, so that they can be combined with or compared to other verification modalities like speech [13]. As an illustration of the usefulness of the discriminant measure over all classes, we show the ROC for the residual matching error (RME) and the local discriminants (LD) in Figure 5. Such curves reflect the performance of a given solution averaged on *all* classes. The points on the ROC were obtained by scaling the minimum threshold displayed in Figure 4 with a varying factor. Clearly, the LD outperforms the RME everywhere. At the threshold value 0.5, the false alarm rate is 6.8% and the false acceptance rate is 3.6%.

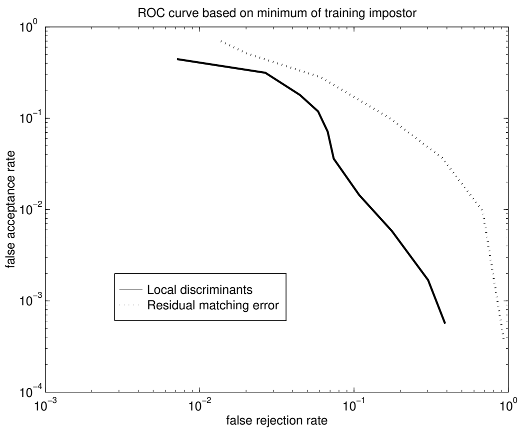

Fig. 5. Experimental ROC curve for the residual matching error and the local discriminants in a log-log scale. Results were obtained with $\lambda = 2$.

Figure 6 shows the LD measures for a particular person, revealing which nodes have little relevance for discrimination. The LD approach provides an automatic way of suppressing the nodes which do not contribute to authentication.

Evaluation of Elasticity Significance In order to assess the effectiveness of grid elasticity, we compare an elastic and a non-elastic graph matching procedure. The non-elastic graph matching is obtained by skipping the second step of the matching procedure described in Section 4.1, which is equivalent to choosing a very large λ in (7). A completely "plastic" grid is obtained with $\lambda = 0$: as the second term vanishes, each grid node is free to move in the image. By running the simulations according to the expert protocol of Section 3.2 with several values of λ, it is possible to assess the usefulness of the elastic step, and also to study the tolerance of the discriminant approach with respect to the rigidity of

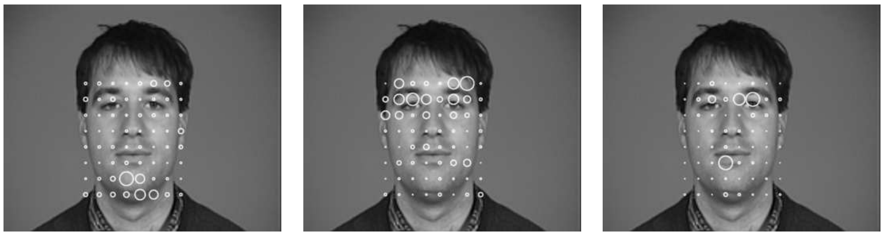

Fig. 6. LD measures at each node, for three resolutions. The lowest resolution is shown left, the highest resolution is shown right.

the grid. To the best of our knowledge no such quantitative analysis of λ has been documented before. For preventing any convergence problems at low values of λ, the number of iterations on the elastic matching was limited to 100.

Figure 7 shows the total error rate defined by TE=FA+FR, for the rigid matching and the elastic graph matching, for both types of discriminant measures. Clearly, the presence of the local discrimination has a larger influence on the results than the elastic deformation.

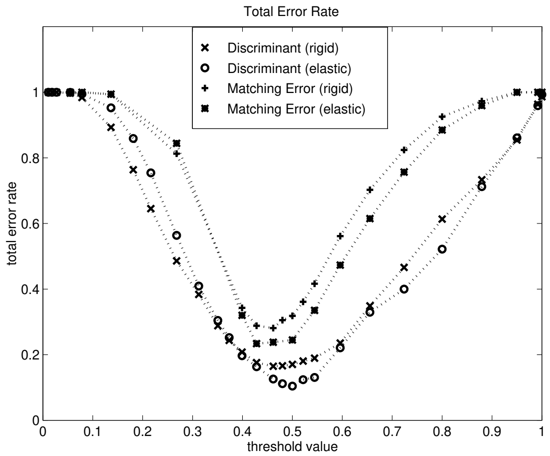

Fig. 7. Total error rates according to the threshold for the rigid matching and the elastic graph matching, with $\lambda = 2$.

Results at the 0.5 threshold are shown for several values of λ in Table 1. The *Equal Error Rate* (EER), defined as the point where FA = FR, is also shown. There is a transition from elastic to rigid matching. The local discrimination is able to provide almost constant results for λ between 0.5 and 3.0. For larger values of λ, the performance degrades. The elastic graph matching improved the rigid graph matching, which can be observed by inspecting Figure 7. Table 1 shows that the EER is improved from 14% down to 11%. However, combining the rigid graph matching with local discriminants is better than elastic graph matching. Not surprisingly, combining the elastic deformation with local discrimination yielded the best results.

λ	**FR**	**FA**	**EER**
∞	29.7	2.1	14.4
10.0	26.3	2.2	12.0
5.0	25	2.3	11.9
4.0	22.9	2.5	12.1
3.0	22.9	2.4	12.3
2.0	22.3	2.2	11.8
1.0	21.6	2.3	11.2
0.5	21.6	2.0	10.8
0.0	24.3	1.6	11.6

λ	**FR**	**FA**	**EER**
∞	13.3	3.7	8.5
10.0	11.1	3.6	7.3
5.0	10.7	3.5	8.4
4.0	11.3	3.6	9.2
3.0	6.9	3.6	6.5
2.0	6.8	3.6	6.1
1.0	7.1	3.9	5.4
0.5	6.8	4.3	6.2
0.0	9.4	4.0	6.0

Table 1. Error rates at the 0.5 threshold and Equal Error Rates. (a) with residual matching error as dissimilarity measure, (b) with local discriminants. The rigid case corresponds to a very large λ, denoted here by $\lambda = \infty$. The equal error rates are obtained by interpolation.

As a conclusion, it has been shown that a small degree of elasticity provides an improvement of the performance. The behaviour remains constant over a certain range of λ, but from a certain rigidity on, the performance degrades.

4.5 Eye Detection by Saccadic Search for Normalisation

It is known that, if face images are normalised, the authentication performance of the matching system is improved. Normalising the ocular positions is such a procedure which can be implemented in an active vision based face authentication. Here we suggest to detect the eye positions of a person by using the Gabor responses dynamically. We use a rigid graph composed of nodes on concentric circles obtained by log polar mapping as in Figure 8. As the procedure is not person specific it can also be used in identification applications. The performance may be of course improved if person specific eye models are used. However, the face expert we described above functioned without eye normalisation on the M2VTS database. The eye normalisation technique suggested below is not intended for a database in which the person is already in the central part of the image and has approximately the correct size, but rather a dynamic environment where the camera is active in order to get the best takes of a face.

Saccadic Search At the beginning of the search, the retinal sampling grid is placed at a random position on the image and the corresponding set of Gabor features at grid nodes, represented by the set $\mathcal{G}_0$, is extracted. Each vector in $\mathcal{G}_0$, after division by its Euclidean norm, is subsequently matched against a reference vector e_{av}. In order to construct the latter, the average Gabor responses from the centre of the right and left eye of six persons are computed. These two standard vector responses are then geometrically averaged component-wise so that e_{av} captures the features which are common to the right and the left eye. The point of the grid for which the Euclidean distance from e_{av} is minimal is selected as the target for the next saccade. The search is terminated when saccades become short, here shorter than 1/6 of the sampling grid's outer radius. If no saccade target whose distance from e_{av} is reasonably low can be found (which can be the case if the search starting point happens to fall in a blank region of the image), the search is restarted from a random position.

The Eye Model The a priori knowledge about the appearance of the left and right eyes of the generic person is respectively encoded into a left eye model and a right eye model. The models are constructed from the sets $\mathcal{L} = \bigcup_p \Gamma_p$ and $\mathcal{R} = \bigcup_q \Gamma_q$ of Gabor features obtained by placing the retinal sampling grid on either of the eyes (Figure 8) and computing the Gabor responses Γ_p at each of its points.

Fig. 8. The retinal sampling grid placed on a person's right eye for model creation.

The features in $\mathcal{L}$ and $\mathcal{R}$ are then rearranged in a collection of matrices $\mathsf{M} = \{\mathsf{M}_{r\omega}\}_{r\omega}$ so that each one of the $\mathsf{M}_{r\omega}$ contains the responses for a fixed Gabor frequency radius ω and a given spatial circle with radius r of the sampling grid. The rows and columns of each $\mathsf{M}_{r\omega}$ therefore correspond to the variation of the angular coordinates in the spatial and frequency domains. Matrices $\mathsf{M}_{r\omega}$ are then normalised separately with respect to the norm defined by $|\mathsf{M}_{r\omega}| = \sqrt{\mathrm{Trace}(\mathsf{M}_{r\omega}^t \mathsf{M}_{r\omega})}$, which is equivalent to the Euclidean norm if $\mathsf{M}_{r\omega}$

is interpreted as a vector. All Gabor features from a single frequency channel ω belong to the same matrix $\mathsf{M}_{r\omega}$. Since each frequency channel is characterised by a specific bandwidth which is common to all the orientations, normalisation takes care of the variation of filter bandwidths across the frequency channels. Also, by grouping together all the points of the sampling grid circle of radius r in a single $\mathsf{M}_{r\omega}$ and then normalising, one makes sure that illumination changes are compensated for.

The eye model for the left eye, L, is computed by combining the collections of matrices M^i obtained by placing the grid manually on the left eye of six persons according to the relation

$$\mathsf{L} = \{\mathsf{L}_{r\omega}\}_{r\omega} = \left\{ \frac{\sum_i \mathsf{M}^i_{r\omega}}{|\sum_i \mathsf{M}^i_{r\omega}|} \right\}_{r\omega}$$

The same procedure is applied to obtain the eye model for the right eye.

Matching of the retinal grid samples I extracted from an image with the model is performed (e.g. in the case of a left eye) by minimising the value of the function $d(\mathsf{I}, \mathsf{L}) = \sum_{r\omega} |\mathsf{I}_{r\omega} - \mathsf{L}_{r\omega}|$.

Refining the Search After the saccadic phase of the search has converged to the target pattern, the Gabor responses in the points currently "viewed" by the grid are compared with both the left and the right eye models described in the preceding section. According to the model which obtains the best result, the candidate eye is assumed to be a left or a right eye. The appropriate model is then selected and the exact position of the local minimum is determined. If the resulting displacement is larger than a few pixels the saccadic search is restarted from a random position.

Experiments have shown that the saccadic search may detect some erroneous local minima (e.g. the corners of the mouth, ear-rings or details in the hair). In order to discriminate such fake targets, the difference is computed between the candidate's distance from the attributed eye model and its distance from the alternate model. The ratio of this difference to the minimum distance, which we call the *asymmetry*, measures the amount to which the chirality of the detected feature contributes to the match. In our experiments, the asymmetry always turned out to be grater than 0.1 for correct matches, while it generally dropped of one or two orders of magnitude in the case of spurious identifications. The errors thus detected are treated by restarting the search from a random position.

Looking for the Other Eye After localisation of one eye, the system performs a saccade in the presumed direction of the other eye. Normal saccadic search is then performed until an eye is found. Due to scale differences between images, the initial saccade may not turn out to be long enough to prevent the system from finding again the same eye. In this case, further attempts are performed with an increasing starting distance from the known eye until the other eye is found. In case the search refinement detects a low asymmetry target, search is restarted with a random offset. If this condition persists for several attempts, it is assumed that the position of the first eye has been incorrectly assigned and eye detection is restarted from scratch. Although the assumption that faces are presented in an upright orientation is used to speed up the detection of the second eye, no strict constraint is imposed on its position relative to the first. Therefore, detection remains robust also in the case of subjects having their head tilted to one side.

Experimental Results The algorithm has been tested using a Gabor decomposition rosette consisting of six texture orientation sectors and five frequency magnitude octaves, ranging from $\frac{\pi}{16}$ to π. The retinal sampling grid employed had 5 rings and 16 rays, with the ring radii being distributed between $\rho_{\min} = 3$ and $\rho_{\max} = 30$ pixels.

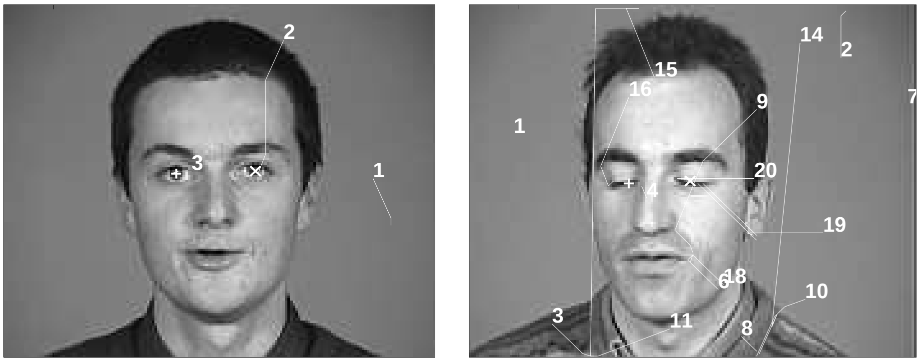

Fig. 9. The + and × signs denote the best match with the right and left eye models respectively. Numbers identify successive starting points for saccades. Eye detection for the left picture required 51 fixations. Note how saccadic search 1 was considered uninteresting and therefore discarded. A random restart (2) then lead to detection of the left eye, after which saccadic search resumed (3) near the location of the right eye. In the case of the right picture, information from the outline of the orbit allows eye detection even if the person's eyes are shut. During this trial the centre of the sampling grid explored 99 pixels and 14 targets were rejected after comparison with the eye models.

Our test set consists of forty takes of twenty persons from the M2VTS database. The image resolution is 143 × 175 pixels. Differences between the takes of the same persons consist in tan changes, haircut, makeup, eyelid position, head position (heads are often slightly rotated) and slight scale changes. Several persons in the database wear eyeglasses.

Single takes from six persons were used to extract the left and the right eye models. Repeated testing was then performed on the whole set without any mismatch being found. Information obtained from the outline of the orbit allows correct detection of the features even when the subject's eyes are closed (Figure 9). In our trials we found the median of the number of fixation points to be 49 for the detection of both eyes, that is to say that the centre of the retinal sampling grid explores 0.2% of the image pixels. The number of fixations is considerably increased (typically 100) for subjects wearing glasses with strong reflections or having their eyes shut. This is mainly due to the fact that since the algorithm knows nothing about facial features other than the eyes, no alternative cues can be used to infer their spatial position when their visibility is low. Nevertheless, detection is always correctly accomplished at the end. However, the results are indicative. The performance of the method should be tested using an active camera setup in the future.

5 Speech Expert

5.1 Feature Extraction

One of the earliest applications of speech features as biometrics is forensics. The physical and behavioural phenomena which help making the speech so personal include, the characteristics of the vocal tract, the shape of the oral cavity, the nerve signals, and muscle dynamics. The interplay and the exact role of the different elements influencing the characteristics of the speech is too complex to be identified through the resulting one dimensional signal, the voice. However, many personal characteristics are possible to capture in the local power spectra of this signal.

The Linear Prediction Coefficients, (LPC) as derived from the Cepstrum information, is the local spectral information which is most frequently utilised in speech processing in general and speaker authentication in particular. The LPCs, their first and second order time derivative approximations (first and second deltas) are commonly used together as a feature vector describing the characteristics of speech, in typically 10 ms of partly overlapping time intervals, [22,17].

5.2 Text Dependent Speaker Authentication

The techniques described here define the second processing step of a speech expert. As our speech expert, an implementation of the work in [18], uses the fusion of decisions coming from three matching algorithms to deliver a final opinion, we present these below. The final combined graded opinion which is obtained by weighting the individual decisions with the distance to the decision threshold (used in decision making of each method) for each client. The LPCs are used as feature vectors in all three methods. The ROC curves of the speech expert alone is given by Figure 10.

Dynamic Time Warping This is a template matching technique which has many similarities with our face authenticator technique in that the reference

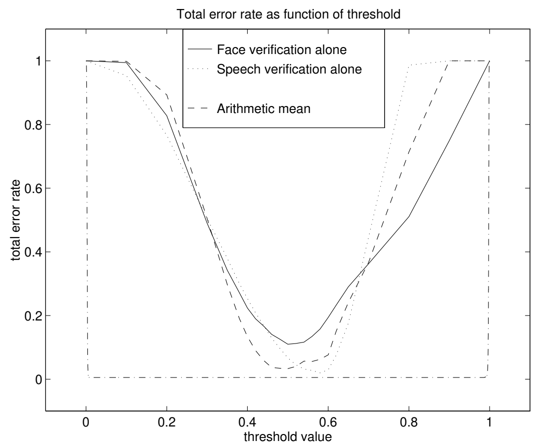

Fig. 10. The total error rates of speech modality as compared to face modality.

feature vector sequence is warped (geometrically distorted) towards the test sequence and a scalar product is performed between the two. The time warping attempts to align the test and the reference speech features in that the changing speed with which the speech is uttered, is normalised and the feature vectors are possible to compare with each other after the warping, [24].

Sphericity The reference LPC sequence x_k defines the covariance matrix

$$X = \frac{1}{M} \sum_{i=1}^{M} x_i x_i^t \tag{11}$$

where M is the total number of the local analysis intervals. Similarly the test LPC sequence covariance matrix, Y is obtained. If the dimension of the X is $m \times m$ the sphericity measure is defined as

$$\mu(X,Y) = \log \frac{m^{-1}\,\mathrm{trace}(YX^{-1})}{m(\mathrm{trace}(XY^{-1}))^{-1}} \tag{12}$$

The larger the sphericity measure between two vectors is, the more likely it is that they represent two different speakers [7].

Hidden Markov Models, HMM HMMs have been used to model the time series. A major use of HMMs is to build models of sub-parts of speech, such as phonemes, or words, see [22] for a tutorial. Here we used text dependent speaker authentication by using the digits $\{0..9\}$. One way of exploiting HMMs in speaker verification consists in creating one set of models for the client and one (small) set of models for all impostors (world model), [23]. The two sets of models contain the HMM models of the digits, as uttered by a client and as uttered by the world. A decision is made by computing

$$\arg \max_{\omega \in \{\mathrm{CL},\ \mathrm{IM}\}} \{P(\omega|O)\} \tag{13}$$

where O is the observed speech (feature vector), and CL and IM represent client and impostor respectively, by using Bayes rule $P(\omega|O) \propto P(O|\omega)$. The latter distribution, $P(O|\omega)$ is modelled by replacing ω with M, a Markov model of the uttered word (by a client or an impostor) ass $P(O|\omega) \propto P(O|M)$. This is in turn modelled as a Markov chain with unknown states (the number of states are known), unknown transition probabilities between the states, and a model of the symbol probability distribution for each state. Computable estimations of $P(O|M)$ are obtained through training which uses the well established Viterbi algorithm and the Baum-Welch re-estimation for doing so. The parameters of the world model is speaker independent. In our case the client set consisted of the speech takes of the M2VTS data-base, [14], whereas the world model was computed by a separate database consisting of 300 occurrences of each digit (uttered by 500 persons), [18]. Furthermore, the number of states of the digits were determined by allocating each phoneme one state, and the model of the symbol probabilities was assumed to consist of one Gaussian per state. All digits had a left-right structure as state transition model.

6 Opinion Fusion by Supervisor

A more extensive presentation of the mathematical background of the model we used can be found in Bigün [3,2].

Basics of the Supervisor Algorithm We perform the following steps

1. (Supervisor Training) Estimate the bias parameters of each expert, i.e. $\{M_i, V_i, \alpha_i\}$, according to (14)

$$M_i = \frac{\sum_{j=1}^{n} \frac{z_{ij}}{\sigma_{ij}^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_{ij}^2}} \quad \text{and} \quad V_i = \frac{1}{\sum_{j}^{n} \frac{1}{\sigma_{ij}^2}} \tag{14}$$

   by using a training set i.e. x_{ij}, y_j, and p_{ij} with j up to n. The bias parameters will be computed for each expert by using all available persons in the training set. σ_{ij}^2 are computed according to (15), (16).

$$\bar{\sigma}_{ij}^2 = \frac{\alpha_i}{p_{ij}^2} = \frac{(G_i - D_i)}{n-3} \cdot \frac{1}{p_{ij}^2} \tag{15}$$

$$G_i = \sum_{j=1}^{n} \left(\frac{z_{ij}^2}{s_{ij}} \right) \quad \text{and} \quad D_i = \left(\sum_{j=1}^{n} \left(\frac{z_{ij}}{s_{ij}} \right) \right)^2 \left(\sum_{j=1}^{n} \left(\frac{1}{s_{ij}} \right) \right)^{-1} \tag{16}$$

2. (Authentication Phase) At this step, the supervisor is operational, meaning that the time instant is always $n+1$ and that the supervisor has access to expert opinions $x_{i,n+1}$, and $p_{i,n+1}$, but not access to the true authentication scores, y_{n+1}. The expert opinions are normalised yieldingM', and V' according to (17).

$$M_i' = x_{i,n+1} + M_i \quad \text{and} \quad V_i' = V_i + \sigma^2_{i,n+1}. \tag{17}$$

M'' and/or V'' are computed according to (18) (and are ready to be thresholded to yield a definite decision).

$$M'' = \frac{\sum_{i=1}^{m} \frac{M_i'}{V_i'}}{\sum_{i=1}^{m} \frac{1}{V_i'}} \qquad \text{and} \qquad V'' = \sum_{i=1}^{m} \frac{1}{V_i'} \tag{18}$$

$\sigma^2_{i,n+1}$'s are computed according to (15).

Score Transformation

Depending on the algorithms they use, the scores of the experts, X_{ij}, may or may not be dimensionless (scaled) or in the correct range i.e. $[-\infty, \infty]$. The prime "$'$" on X and Y variables represent these variables before transformation. For our purposes, the transformation

$$X_{ij} = \log \frac{X_{ij}\prime}{1 - X_{ij}'} \tag{19}$$

which is also known as the "odds of X_{ij}'", will be used to map the scores in $[0,1]$ to $[-\infty, \infty]$.

Fusion Experiments In Table 2 we present the minimum total error rates of the speech and the face modalities individually, the Bayesian supervisor, and the plain mean of the scores of the face and speech experts, as an alternative supervisor. The test followed the Supervisor Test Protocol described earlier. The FA and FR curves of the Bayesian Supervisor are much smaller than those corresponding to the Mean Supervisor, Figure 11. Furthermore, the minimum total error rate for the Bayesian supervisor is **0.006** , which should be compared to that of the Mean Supervisor, **0.015** . However, in both cases there is a significant improvement as compared to individual modalities, Table 2. While the standard threshold yields the lowest TE for the Bayesian supervisor, this figure is increases to 0.0165 for the Mean Supervisor.

These and other experiments indicate that the Bayesian supervisor is more successful in decision making due its capability of symmetrising the score error densities.

TE_1	TE_2	TE_{bs}	TE_{ms}
0.056	0.035	0.006	0.015

Table 2. Minimum total error rates of machine supervisor opinions based on face and speech signals. TE_1 and TE_2 are expert minimum total error rates of face and speech respectively.

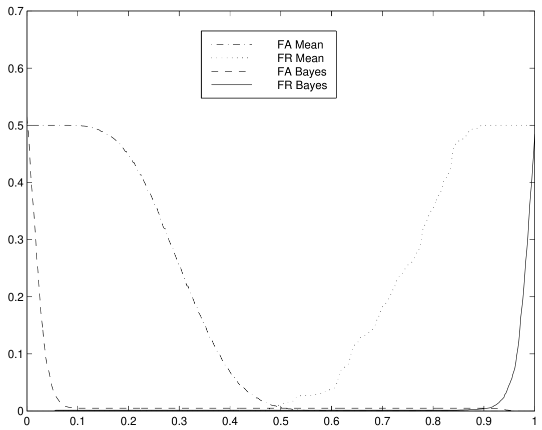

Fig. 11. False Acceptance and False Rejection curves for Mean and Bayesian supervisor tested on 1 speech and 1 image expert.

7 Conclusions

We have presented a framework for multi-modal person authentication, this included test procedures using a bootstraping technique, modelling the experts and the supervisor as opinion providers rather than hard decision makers. We implemented a face expert based on Gabor decomposition, a speech expert using LPCs, and a supervisor based Bayesian statistics and evaluated the individual experts as well as the supervisor on real data.

We demonstrated that a multi-modal system is capable of improving decisions in the context of person authentication significantly, by decreasing the total error rate as much as 600 % (reaching the rate of 0.006 on a rotational test procedure) as compared to the best modality.

In addition to the the general framework, our contribution has been in i) improving the Elastic Graph Matching approach by Local Discriminants, ii) quantifying the contribution of the elastic part of the matching as compared to the rigid graph matching, iii) proposing log-polar based eye detection by saccadic movements for image normalisations for a dynamic camera, and iv) proposing the Bayesian Supervisor in order to improve the multi-modal decision making.

Acknowledgements

This work has been supported by the European projects IT-VIRSBS and ACTS-M2VTS. We thank Gilbert Maitre, IDIAP, Switzerland, for his help with speech experiments.

References

1. J. M. Bernardo and M. F. A. Smith. *Bayesian Theory.* Wiley and Son, Chichester, 1994.
2. E. S. Bigün. Risk analysis of catastrophes using experts' judgements: An empirical study on risk analysis of major civil aircraft accidents in Europe. *European J. Operational research*, 87:599–612, 1995.
3. 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 311–318. Springer, 1997.
4. J. Bigün. Unsupervised feature reduction in image segmentation by local transforms. *Pattern Recognition Letters*, 14:573–583, 1993.
5. J. Bigun, G. Chollet, and G. Borgefors; Eds. *Proceedings of the first international conference on Audio and Video based Person Authentication - AVBPA97*, volume LNCS-1206. Springer, 1997.
6. J. Bigun and J. M. H. du Buf. N-folded symmetries by complex moments in Gabor space. *IEEE-PAMI*, 16(1):80–87, 1994.
7. F. Bimbot and L. Mathan. Second order statistical measures for text independent speaker identification. In *ESCA*, pages 51–54, 1994.
8. R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):955–966, October 1995.
9. P. J. Burt. Fast filter transforms for image processing. *Computer Graphics and Image Processing*, 16:20–51, 1981.
10. J. G. Daugman. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, July 1988.
11. P.A. Devijver and J. Kittler. *Pattern Recognition: a Statistical Approach.* Prentice-Hall International, London, 1982.
12. U. Dieckmann, P. Plankensteiner, R. Schamburger, B. Froeba, and S. Meller. SESAM: A biometric person identification system using sensor fusion. In J. Bigun, G. Chollet, and G. Borgefors, editors, *Audio and Video based Person Authentication - AVBPA97*, volume LNCS-1206, pages 301–310. IAPR, Springer, 1997.
13. B. Duc, G. Maître, S. Fischer, and J. Bigün. Person authentication by fusing face and speech information. In J. Bigün, G. Chollet, and G. Borgefors, editors, *First International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA'97)*, volume 1206 of *LNCS*, pages 311–318, Crans-Montana, Switzerland, March 12-14 1997. Springer.
14. B. Duc, G. Maître, S. Fischer, and J. Bigun. Person authentication by fusing face and speech information. In J. Bigun, G. Chollet, and G. Borgefors, editors, *Audio and Video based Person Authentication - AVBPA97*, volume LNCS-1206, pages 311–318. Springer, 1997.

15. F. Dufaux and F. Moscheni. Motion estimation techniques for digital TV: A review and a new contribution. *IEEE Proceedings*, 83(6):858–876, June 1995.
16. B. Efron and R. J. Tibshirani. *An introduction to the Bootstrap*. Chapman & Hall, New York, 1993.
17. S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech, Signal Processing*, 29(2):254–272, 1981.
18. D. Genoud, F. Bimbot, G. Gravier, and G. Chollet. Combining methods to improve speaker verification decision. In *Proceedings of The Fourth International Conference on Spoken Language Processing*, Philadelphia, October 3-6 1996. ICSLP.
19. P. Jourlin, J. Luettin, D. Genoud, and H. Wassner. Acoustic-labial speaker verification. In J. Bigun, G. Chollet, and G. Borgefors, editors, *Audio and Video based Person Authentication - AVBPA97*, volume LNCS-1206, pages 319–326. IAPR, Springer, 1997.
20. M. Lades, J. Buhmann J. C. Vorbrüggen, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, March 1993.
21. S. Pigeon and L. Vandendorpe. The M2VTS multi modal face database (release 1.0). In J. Bigun, G. Chollet, and G. Borgefors, editors, *Audio and Video based Person Authentication - AVBPA97*, pages 403–409. Springer, 1997.
22. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recogniton. *Proceedings of the IEEE*, 77(2):257–286, 1989.
23. A. E. Rosenberg, C. H. Lee, and S. Gokon. Connected word talker verification using whole word hidden Markov model. In *ICASSP*, pages 381–384, 1991.
24. H. Sakoe and Chiba. Dynamic programing algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech, Signal Processing*, 26(1):43–49, 1978.