A fully automatic face recognition system using a combined audio-visual approach [*]

Alberto Albiol[†], Luis Torres[†], and Edward J. Delp[⋆] [†]
[†]Communications Department
Technical University of Valencia, Valencia, Spain
alalbiol@dcom.upv.es
[†]Department of Signal Theory & Communications
Technical University of Catalonia, Barcelona, Spain
luis@gps.tsc.upc.es
[⋆]School of Electrical and Computer Engineering
Purdue University West Lafayette, IN 47907-1285
ace@ecn.purdue.edu

Corresponding Author:
Dr. Alberto Albiol
Communications Department
Technical University of Valencia, Valencia, Spain
46022 Valencia (Spain)
Telephone: +34 96 387 97 38
Fax: +34 96 387 73 09
Email: alalbiol@dcom.upv.es

Abstract

This paper presents a novel audio and video information fusion approach that greatly improves automatic recognition of people in video sequences. To that end, audio and video information is first used independently to obtain confidence values that indicate the likelihood that a specific person appears in a video shot. Finally, a post-classifier is applied to fuse audio and visual confidence values. The system has been tested on several news sequences and the results indicate that a significant improvement in the recognition rate can be achieved when both modalities are used together.

[*] Please address all correspondence relative to this manuscript to Dr. Alberto Albiol.

[†] ACKNOWLEDGMENT: The work presented was partially developed within VISNET,a European Network of Excellence (http://www.visnet-noe.org)

1 Introduction

Automatic machine recognition of people is a challenging problem which has been addressed by many researchers due to its many different applications (see [1] for a list).

Recognition is usually achieved by using one or more biometric keys that identify a person within a population. In most cases, the choice of the particular biometric deeply relies on the final application. For instance, retinal scans and fingerprints have shown high recognition accuracy, however their use is limited to the availability of cooperative individuals, which is not always possible.

Although much attention has been placed on biometric development for security and physical access applications, the recognition of people in video sequences for video indexing applications is also an immediate need with significant commercial opportunity [2]. In this application only face and voice biometrics are usually available. Recognition performance using these biometrics has reached a relatively high success in applications where some control on the acquisition conditions is possible, for instance with controlled pose and illumination settings for face recognition. However, in uncontrolled environments with changes in pose and illumination, face recognition is still an open problem, and it seems that this situation will still remain for some years on. A similar context is also found when recognition is performed using voice-based biometric.

As mentioned before, a possibility to increase the accuracy of recognition systems is to use more than one biometric. This is because degradations for each modality usually are uncorrelated.

Following the previous ideas, this paper presents a system that recognizes people in video sequences combining audio and image information. More specifically we are interested in locating shots where some particular person appears in the image while talking, so that both face and voice biometrics can be used. Examples of these shots include taped footage of news anchors, and head and shoulders sequences of people being interviewed. Notice that this type of shots usually contains important and reusable information, and therefore they are important for indexing. Moreover recording conditions for this type of shots are usually more controlled, making the recognition task more accurate.

Recognition tasks have been traditionally grouped into identification and verification [1, 3, 4] irrespectively of the specific biometric to be used. In identification the objective is to obtain the identity of a test sample from a set of known possible identities. In verification an identity for the test sample is proposed and the task entails accepting or rejecting this identity claim. Usually, during the identification task the test sample is compared with all possible identities and it is assigned to the one with highest similarity. In the case of verification, the test sample is compared to the claimed identity model and usually a threshold value is needed to make the decision. The use of this threshold makes the verification task harder than the identification task, and this can explain why most research on person recognition has concentrated on the identification task [4, 5, 6].

In this paper we will focus on the verification task. Thus, if person m is being searched, then, for each test shot, the identity m will be claimed and the system will accept/reject this identity claim. Following with the terminology used in verification systems, if person m was really in the shot we say that it was a *true claimant* shot. On the other side, if person m was not in the shot then we say that it was an *impostor* shot.

Figure 1 illustrates the approach used in this paper. As shown in the figure, audio and image information are processed in parallel and two confidence values, that express the belief that person m is in the shot, are extracted for each selected shot. This is done using two modules that use face and voice biometrics. Finally, a *fusion* module is used to make the final decision based on the audio and image confidence values. Notice that the previous approach can be applied to every shot within a news sequence for instance. However, since we are interested in recognition in head and shoulders shots, it seems reasonable to filter out as many shots as possible and concentrate recognition only in shots where recognition is possible, before the overall system is applied. An example of a system used to pre-select suitable shots for recognition in news sequences can be found in [7].

The rest of the paper is organized as follows. Section 2 and 3 describe how image and audio confidences are extracted respectively. Then, Section 4 explains how these confidence values are used to accept or reject the identity claim. Finally, some results and conclusions are presented in Section 5.

2 Extracting face confidence

As shown in Figure 1 our approach to extract the face confidence is based on face recognition. Face recognition has been an active research topic for more than one decade. Initially, face recognition systems focused on still images, although, in recent years, face recognition in image sequences has gained significant attention [1]. Image sequences offer the advantage of selecting frames more suitable to the recognition task, although nearly all systems apply still-image face recognition techniques to individual frames. A face tracking approach may also be used for image sequence face recognition. However, face tracking poses some extra problems: need to refine the tracking results to obtain good candidates for recognition, and the computational load associated to the tracking process itself. For these reasons, we have chosen a simpler approach that is based on still-image face recognition applied to individual frames.

Still-image face recognition methods have been compared in different survey papers [1, 8]. The basic conclusion obtained is that although all the algorithms have been successfully used for face recognition, each of them have their own advantages and disadvantages. Thus the appropriate scheme has to be chosen based on the final application. For instance, elastic graph matching techniques [9], require large face resolutions (128 × 128). Other methods are better suited for identification applications, such as Linear Discriminant Analysis (LDA) [10]. In other cases the difficulty of training the face model limits the use of some algorithms as in the case of the Hidden Markov Model (HMM) algorithms [11]. In any case, the comparison results presented in [4], over twelve different face recognition techniques used for verification applications indicate that Elastic Graph Matching, LDA and Principal Component Analysis (PCA) were on the top three, each method showing different levels of performance on different subsets of images. In this paper, face recognition is based on a variant of the PCA technique [12], also known as the self-eigenfaces technique [13]. The self-eigenfaces technique is well suited for the video indexing applications where many images of a specific person viewed from a similar perspective are available for training purposes. Details about the self-eigenface approach are presented in Section 2.1.

Figure 2 illustrates the main blocks used in this paper to obtain the face confidence. The input of the system is the shot where we want to recognize person m. Then, for each frame, still-image face detection/recognition is applied. As mentioned in the previous section, any verification task usually entails two steps: comparison against a suitable model and thresholding in order to accept/reject the identity claim. The idea here is to skip the second step, so that no hard decisions are taken at any point in this stage. The recognition results obtained for each frame are finally combined to yield a single face confidence value per shot.

Traditionally, face detection and recognition have been treated as two separate problems, and it is quite common to find many face recognition systems that assume that the face has been previously manually extracted or detected with independent face detection systems. In our proposal, the key objective of the face detection module is to provide good candidates for recognition. For this reason the details of face recognition are explained first in Section 2.1 to better understand what type of face candidates are needed, and then, the algorithm used to extract these face candidates is described in Section 2.2.

2.1 Face recognition using self-eigenfaces

PCA has been one of the most successful approaches used for face recognition. The PCA technique has been mostly used for identification applications where usually just one frontal face is available for each class [12].

The basic idea behind PCA analysis is to achieve dimensionality reduction by projecting the face pattern onto a subspace of a much lower dimension than the number of points of the face pattern itself. This subspace is spanned by a set of orthonormal vectors usually referred to as *eigenfaces* and it is designed to optimally represent the distribution of face data in the RMS sense. For each application the eigenfaces are obtained from the set of available training faces (usually one face per person). Finally, identification is done by locating the image in the database whose projection coefficients are closest to the projection coefficients of the test image.

In this paper, a variant of the PCA approach which is more suitable for applications where many training faces per person are available is used [13]. The main difference is that independent PCA is performed for each person using his/her available training faces. The result of this analysis is a set of eigenfaces for each person, called *self-eigenfaces*. Therefore each person is modeled by the set of parameters:

$$\mathcal{F}_m = \{\mathbf{x}_\mu^m, \mathbf{V}_m\} \tag{1}$$

where $\mathbf{x}_\mu^m$ is the average face of person m and $\mathbf{V}_m$ is a matrix whose columns are the K_m principal self-eigenfaces $\mathbf{v}_k^m$ of person m. To achieve the dimensionality reduction K_m is much smaller than the number of pixels of the face patterns. In our approach, the location of the eyes is used to normalize the size of each training face. Figure 3.a presents some original views of a person and Figure 3.b an example where the average face and the self-eigenfaces corresponding to the largest eigenvalues of the same person are shown.

The test stage in the self-eigenface approach also differs from the original PCA approach. As mentioned previously, in the original approach projection coefficients are used for classification. In the self-eigenface approach each face candidate is projected and reconstructed using a particular set of self-eigenfaces $\mathcal{F}_m$. Finally, the reconstruction error is used to measure the confidence such that the identity of the face candidate is m.

Let $\mathbf{x}$ be a test face represented as a column vector, its projection coefficients using the self-eigenfaces of person m can be obtained as:

$$\mathbf{y}^m = \mathbf{V}_m^T(\mathbf{x} - \mathbf{x}_\mu^m) \tag{2}$$

Using the projection coefficients $\mathbf{y}^m$ the original test pattern can be approximated as:

$$\tilde{\mathbf{x}}^m = \mathbf{V}_m.\mathbf{y}^m + \mathbf{x}_\mu^m \tag{3}$$

Once the test image has been reconstructed, the reconstruction error is evaluated as:

$$\epsilon = \frac{1}{255}\sqrt{\frac{1}{RC}\sum_{j=1}^{RC}|\mathbf{x}(j) - \tilde{\mathbf{x}}^m(j)|^2} \tag{4}$$

where j is the vector index, and R and C are the number of image rows and columns respectively.

The idea behind this method is that given a test face, a low reconstruction error ϵ (good fit) is achieved when the self-eigenface set of the corresponding identity is used. Figure 4 illustrates this idea where original and reconstructed test faces are shown. In both cases, the faces are reconstructed using Equation 3 with the mean face and self-eigenfaces of Figure 3.b. It can be seen that reconstruction error is smaller when the identity of self-eigenfaces matches the identity of the test face.

The self-eigenface technique can be easily extended to video sequences by repeatedly applying the face recognition to every frame and then, giving a global confidence value that person m appears in the sequence. A practical way to obtain a global confidence measurement, can be done using the median value: Let ϵ_i be the minimum reconstruction error for all the face candidates of frame i. Then, we define:

$$FC_m = \text{median}\{\epsilon_0, \epsilon_1, \epsilon_2, \ldots, \epsilon_N - 1\} \tag{5}$$

as the shot-based confidence that the face of the person m appears in a particular shot. One advantage of the median value compared to other global measurements, such as the mean value, is that it is more robust to outliers.

If only visual information is available, simple thresholding of FC_m can be used for verification.

In general, we can say that the self-eigenface approach works well as long as the image under test is similar to the ensemble of images used in the calculation of the self-eigenfaces, because only linear combinations of the self-eigenfaces are used to reconstruct the test face. This also implies that the facial features in both training and test faces have to be in the same locations, so that it is possible to reconstruct them using linear combinations.

2.2 Face detection

One key point for any face recognition system is face detection. Face detection is also important in other fields such as facial expression recognition and surveillance. For these reasons many different approaches have been proposed in the recent years to address this problem (see [14] for a survey).

A common issue of many face detectors found in the literature is that no specific application is proposed for the detected faces (although face detection itself might be an interesting application). Therefore it can be the case of correct detected faces might not be useful for face recognition. For this reason we prefer to design a simple face detection stage and let the face recognition stage be part of detection stage. Another reason for using a simpler approach is that the objective is to extract faces from head and shoulders sequences, which can be considered as an *easy* case for most face detectors (frontal view in upright position). Notice that the face detection scheme proposed below is not very adequate to detect face rotations or pose variations and will lead to some missing faces in these circumstances. However, in this application, this is not a very important problem, as the face recognition stage would not be able to recognize faces with strong pose variations either.

As in other approaches, our algorithm exhaustively examines all pixel locations at different image resolutions using a fixed size window [15]. This multi-resolution step allows to deal with changes in scale. The objective is to obtain a distance to a face model for every pixel and image resolution obtaining a set of what we call *distance images*. Face candidates are extracted at the local minima of the distance images.

As described in Section 2.1, if we want to recognize a face, the location of its facial features must match those of the training faces. Therefore, the distance to the face model used in the face detection stage is based on the location of facial features, so that good candidates for recognition can be obtained. Figure 5 illustrates how facial features are extracted from a set of sample faces. First, a zero mean Laplacian filter is applied in order to enhance face features, and then a morphological erosion with a 3×3 square structuring element is used to remove white areas from eyes and tooth.

Secondly, to eliminate non-useful features, an intensity thresholding is applied. The threshold is set so that 20% of dark pixels are under the threshold. This setting stems from the observation that facial features approximately occupy this area for frontal faces. The face model $\mathbf{F}_{av}$ used to obtain the distance at each pixel was obtained by averaging 280 patterns as the ones shown in Figure 5.c.

The obtained face model is shown in Figure 6. Notice that the facial features of all images used to obtain the face model are located approximately in the same position as those used to obtain the self-eigenfaces for each person and that a elliptical mask is applied to reduce background noise. Let $\mathcal{W}_p$ be the pixels under the window at a pixel location p after intensity normalization and thresholding (as in Figure 5.c) then the distance to the face model $\mathbf{F}_{av}$ is obtained as:

$$D(\mathbf{F}_{av}, \mathcal{W}_p) = ||\mathbf{F}_{av} - \mathcal{W}_p|| \tag{6}$$

In order to reduce the computational burden, the distance $D(\mathbf{F}_{av}, \mathcal{W}_p)$ is only computed if two simple test are passed. The first test uses skin color to reduce the search of face candidates to skin-colored areas. The second test aims to discard flat skin-colored areas. This is done analyzing the histogram of the intensity values of the pixels under the window.

Compared to other face detection schemes, our approach yields a higher false alarm for a similar detection rate [15]. This is explained for the simplicity of the distance $D(\mathbf{F}_{av}, \mathcal{W}_p)$. It is not uncommon to obtain 5-10 face candidates when just one frontal face is in the scene. However, since the final goal is recognition and not detection, this is not a real drawback of our approach. The advantage is that usually one of those face candidates will be very good for recognition (due to the location of facial features), and the rest of wrong candidates will yield a high reconstruction error after using Equation 4, therefore since the minimum value is taken to obtain ϵ_i in Equation 5 they will be discarded in the recognition stage. In some sense, the face detection stage uses the face recognition stage to discard these erroneous face candidates avoiding the use of a more complex classifier.

3 Extracting voice confidence

Person recognition techniques that use the voice as a biometric are usually referred to as speaker recognition techniques. Note that the objective here is not to know *what* is being said (speech recognition) but *who* says it. Speaker recognition techniques are frequently grouped into *text-dependent* and *text-independent*. In the first case the spoken sentence in both training and testing stages has to be the same. This requirement can be accomplished in some applications such us security access, with the advantage of a higher recognition performance. However, in our application no assumptions about the contents of the speech are made and, then, text-independent speaker recognition techniques must be used.

Before speaker recognition can be performed it is necessary to extract speaker dependent features from the audio signal. Most of the features used for speaker recognition are obtained from the speech spectrum. This is because the spectrum directly reflects the effect of the person's vocal track, which is the main physiological factor to distinguish the voice's identity. Features extracted from filter-banks are commonly used to characterize the speech spectrum, since they have been found to be more robust to noise [16]. Therefore, cepstral coefficients derived from a mel-frequency filter-bank are used in this paper to represent the speech spectrum. In particular, the first twelve cepstral vector coefficients and their corresponding deltas extracted each 17 ms. using a 34 ms. Hamming window are used. We also remove the silent frames and use cepstral mean subtraction (CMS) [17] to reduce linear channel distortions.

Once speaker dependent features are extracted, classification is usually formulated as a classical hypothesis test based on the log-likelihood ratio:

$$AC_m = \log \left\{ \frac{p(S/m)}{p(S/BM)} \right\} \tag{7}$$

where S is the set of speaker dependent features extracted from the audio signal and $p(S/m)$ and $p(S/BM)$ are the conditional probability density functions of the person m and the background BM respectively [18]. The background model characterizes the hypothesis that the audio was not spoken by person m. The use of this background model is important to normalize the values given by $p(S/m)$. The value AC_m indicates the *confidence* that the speech segment was spoken by person m. As in the case of the face confidence, simple thresholding AC_m can be used for classification when only audio information is available.

Different models can be used to represent the probability density functions used in Eq. 7. For instance Hidden Markov Models (HMMs) [19] have been successfully used in text-dependent speaker recognition systems due to their ability to model temporal variations in the speech signal. However, in text-independent speaker recognition, where no assumptions about temporal variations can be made, Gaussian Mixture Models (GMMs) [20] are preferred. Notice that a GMM can be also regarded as a degenerated HMM with just one state. The main reason why GMMs are preferred is its ability to represent arbitrary densities, in this case of speaker dependent features.

During the training stage, GMM models are built for each person m using 2–3 min of clear speech. The parameters of the GMM models are estimated using the Expectation Maximization (EM) algorithm [21]. On the other hand, the background model is built from 1 hour of speech recorded from a variety of speakers extracted from our video database [22]. To make the background model as universal as possible, special care on the composition of the speaker's universe has to be taken. This can be done by balancing as much as possible the number of male/female utterances and different recording conditions.

After training, each GMM consists of a set of parameters that comprises the mean, covariance matrix and mixture weight of each Gaussian component. This set of parameters for speaker m can be written using the following notation:

$$\mathcal{S}_m = \{p_i^m, \overrightarrow{\mu}_i^m, \Sigma_i^m\} \quad i = 1, \ldots, M \tag{8}$$

where M is the number of Gaussian mixtures used in the model.

Depending on the choice of the covariance matrices Σ_i^m, GMMs can adopt several forms. The model can have one covariance matrix per Gaussian component (nodal covariance), one covariance matrix for all Gaussian components in a speaker model (gran covariance), or a single covariance matrix shared by all speakers models (global covariance). The covariance matrix can also be full or diagonal. Among all these options, GMMs that use nodal diagonal covariance matrix have proved to provide better identification results [23]. Similarly, the background model is built using GMMs. However, in this case, the number of Gaussian mixtures is higher to represent a broader set of possible acoustic classes. In this work, the background model uses 256 Gaussian components. This value is based on a previous analysis presented in [20].

In the test stage we use AC_m as a shot-based confidence measurement that the utterance was spoken by m.

4 Combining audio and face confidences

Systems that combine different information sources such as visual or audio are commonly referred to as multi-modal, where each modality corresponds to a different information source. Recently, the research on many different areas such as video indexing and retrieval [24, 25], and video segmentation [26], is also focusing on developing multi-modal systems that aim to improve the performance of each individual modality. An additional benefit of adopting a multi-modal approach is that system design and implementation complexity can be reduced by using several and complimentary modules instead of just one very sophisticated.

The techniques used to fuse different information sources can be broadly classified into pre-mapping and post-mapping techniques [27]. In the first group information is commonly fused by concatenating features from all modalities into a large feature vector, and then, a single classifier is used to perform the recognition task. Pre-mapping techniques have been mostly used in speech-reading [28] and video segmentation [26], although pre-mapping techniques have also been used for person recognition [29].

Post-mapping techniques combine information after features have been processed using a suitable classifier or *expert* for each modality. An expert is essentially a classifier that provides a soft-decision or confidence value instead of a binary decision. Post-mapping techniques have two main advantages. First, features from the different modalities do not need to be synchronized. Second, post-mapping schemes are more robust in situations where features from different modalities are slightly correlated.

Post-mapping techniques that use experts instead of classifiers are usually preferred since they preserve the information about the confidence on each modality. Exceptions to this rule of thumb can be found in [30, 31].

Techniques used to combine confidence values from different experts include: weighted summation [32, 33, 34], weighted product [35] and post-classifier [6]. In this work we selected a post-classifier approach to combine the face FC_m and voice AC_m confidences. The advantage of the post-classifier option is that it is possible to combine confidence values from different experts, even if their outputs fall in different ranges. This happens because post-classifiers directly map the input values from the confidence space to the decision space. In our application a post-classifier allows a direct combination of a face confidence in terms of a reconstruction error with a voice confidence expressed as a likelihood ratio.

Figure 7 shows the scatter plot of the two-dimensional feature vectors $C_m = (FC_m, AC_m)$, where it can be clearly seen that true and false candidates are better separated in the two-dimensional space. To that end, Bayesian post-classifiers are used in this paper [36]. Bayesian classifiers, are based on the likelihood ratio of the conditional probability density functions (p.d.f) $p(C_m|tc)$ and $p(C_m|im)$. Where (tc) and (im) stand for the true claimant and impostor classes respectively. As in Section 3, the conditional p.d.f's can be modeled using GMMs for their ability to model arbitrary densities. Recognition results using different number of Gaussian Mixtures are presented in Section 5.

5 Recognition results

5.1 The Experimental Data Set

All the experiments described in this paper have been done using video data from the *ViBE* database [22]. The *ViBE* database has been created in Purdue University and contains more than 100 hours of copyright cleared MPEG-1 sequences, which were recorded from miscellaneous television programs. The sequences were digitized at a rate of 1.5 Mbits/sec in SIF format (352×240).

Recognition experiments were conducted in head and shoulders shots only. This was done to avoid shots with nobody present and therefore easy to reject. The head and shoulders shots used in the experiments come from 29 different news programs. All the shots within this set are at least 10 seconds long to obtain reliable voice confidences.

The number of different people that appear in the shots is 70, although only 10 people appear in more than 20 different shots. We obtained the face and voice models of these people that appear more frequently only. This was done for a practical reason, think that usually head and shoulders shots in TV news stories do not last more than 15 seconds, and we need at least 2-3 minutes of audio (8-12 shots/person) to train each speaker model and we also need more appearances (5-10 shots/person) for testing. Hence, the dataset was divided in two groups. One for training the face and voice models, and another for testing.

The recognition experiments are made in the following way. For each test shot one of the ten possible identities is proposed. If the proposed identity and that of the test shot match, we say that the shot is a *true claimant* shot and in other case we say that it is an *impostor*. If the system makes the right decision for a true claimant shot then we say that it is a *true positive*. On the other hand, if the system makes the right decision when an impostor shot is tested, then we say that this is a *true negative*. As mentioned in Section 1, a threshold is usually used in verification problems to make the final decision. Depending on the application, it is possible to change the number of true positives and negatives adjusting the threshold value. However, if we want to compare different recognition systems, it is common to set the threshold to the point where the percentage of true positives equals the percentage of true negatives. We call that point the Equal Correct Rate (ECR).

5.2 Recognition results using a single modality

As mentioned at the end of Section 2.1 if either face or voice modalities are used, verification can be done thresholding the available confidence value. This section contains recognition results using only one modality, these results will be useful to compare the benefit of using a multi-modal approach.

5.2.1 Visual recognition results

In Equation 1 we indicated that the face model $\mathcal{F}_m$ was made of the average face and the K_m main self-eigenfaces. As in the original PCA method, K_m is set so that the self-eigenfaces retain a certain percentage of the training set variance. Table 1 shows the number of self-eigenfaces for each identity as the retained variance changes. It can be seen how model complexity increases with the retained variance. The recognition results shown in Table 2 show that there is a trade off for the retained variance (the best recognition results are obtained for a retained variance of approximately 90%). The idea is that the discrimination power gets reduced if too few self-eigenfaces (it is not possible to reconstruct any test sample) or too many self-eigenfaces are used (all test samples yield small reconstruction errors). Figure 8 shows the true positives and true negatives curves as the threshold on FC_m changes when a 90% of the variance is retained. The ECR point correspond to the point where both curves intersect.

5.2.2 Audio recognition results

As shown in Equation 8, speaker dependent features $\mathcal{S}_m$ are modeled using GMMs. Each of the M Gaussian components in a GMM can be regarded as an acoustic class. The number of components M is a trade off between a coarse approximation and over-fitting of the probability density function, for low and high values of M respectively. In this work we performed recognition results for different values of M. The ECR results are presented in Table 2. It can be seen that the best results are obtained for $M = 32$. A similar conclusion was obtained in a previous paper by Reynolds [23]. Figure 8 shows the true positives and true negatives curves as the threshold on SC_m changes when a 32 GMM is used. Again the ECR point corresponds to the intersection of both curves.

5.3 Audiovisual recognition results

In Section 1 it was explained that a multi-modal approach to the recognition problem can improve recognition results if each modality contributes with new and complimentary information. A simple test to analyze how much information is introduced can be done studying the correlation between the face and voice confidence values. To that end we split the confidence values in two different sets; one for the true claimants and another for the impostor shots:

$$\begin{array}{rl} \mathcal{S}_{tc} \equiv & \{C_m = (FC_m, SC_m) \mid C_m \text{ is true claimant}\} \\ \mathcal{S}_{im} \equiv & \{C_m = (FC_m, SC_m) \mid \ C_m \text{ is impostor}\} \end{array} \tag{9}$$

Then, for each set we obtained the correlation factors:

$$\begin{array}{rl} \rho_{tc} = & \text{corr}(FC_m, SC_m) \mid C_m \in \mathcal{S}_{tc} \\ \rho_{im} = & \text{corr}(FC_m, SC_m) \mid C_m \in \mathcal{S}_{im} \end{array} \tag{10}$$

yielding the following values: $\rho_{tc} = 0.03$ and $\rho_{im} = 0.02$. The previous results show that the overlapping of information of both modalities is very small.

An important problem that we came across when obtaining the recognition results is the relatively small amount of available data to train and test the post-classifiers. This is specially critical for the true claimant samples. A possibility to train and evaluate the performance of the post-classifiers is to use a m-fold cross-validation approach [36]. Using this approach, the training data is split randomly in m subsets (in our experiments $m = 5$). Then, m different training and test experiments are performed. In each experiment, one subset is used for testing and the rest are used to train the post classifier. Then, the average performance of the m experiments is used as an estimate of the post-classifier performance.

As indicated in Section 3 there are several options on the type of covariance matrices when building GMMs. In this work only diagonal covariance matrices have been used to model GMMs. This decision is based on the low correlation between face and voice confidences. Moreover, using diagonal instead of full covariance matrices reduces the number of parameters allowing more reliable estimates. Another issue when building GMMs is the number of Gaussian mixtures. Table 3 shows the ECR results using different number of Gaussian mixtures. The table also presents another result for the particular case where both probability density functions are modeled using normal distributions with equal covariance matrix. This special case was used since it produces a linear decision boundary [36]. The ECR results show that in all cases the multi-modal approach outperforms the recognition results for each independent modality. Differences between the Bayesian post-classifiers are not really significant, although a slightly better performance is reached for the GMM-1 model. In this case the number of parameters to estimate is smaller, producing more reliable estimates. Another interesting point is that using equal diagonal covariance matrices for each class (linear decision boundary), also produces a very good separation between true claimant and impostor classes despite its simplicity. In our opinion this is the best option for the post-classifier, this opinion is founded on the fact that generalization is better for simpler classifiers.

Figures 9, show the decision boundaries for the previous Bayesian post-classifiers. It should be mentioned that these decision boundaries have been obtained using all the available data. It can be seen that differences between the decision boundaries when the number of Gaussian mixtures is increased are not really important. Although it can be seen a tendency to over-fitting as the number of mixtures grows. Figure 10 shows the decision boundaries obtained by using visual only, audio only or audio-visual informations. In the figure is also clearly shown how true claimant and impostor shots are better separated when both information sources are used.

6 Conclusions

This paper has presented a person recognition system that uses audio and visual information. The approach used to fuse both information sources is based on a post-classifier that uses voice and face confidences to recognize a specific person in a video shot. The multi-modal recognition results clearly outperform those obtained using a single modality. One drawback of the proposed system is that it can be fooled by dubbed footage. This is a common situation found in news programs when people use foreign languages. A possible solution to this problem is to disable the audio modality on those cases where we may expect a dubbed audio.

As future work, more modalities such as subtitles (when available) should be incorporated into the proposed framework. Also, other face recognition techniques, more robust to changes in pose and illumination, should be explored to extend the applications to more general environment conditions.

References

[1] W. Zhao, R. Chellappa, J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, December 2003.

[2] N. Dimitrova, H. Zhang, B. Shahraray, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," *ACM Computing Surveys*, vol. 9, no. 3, pp. 42–55, Jul-Sept 2002.

[3] G. R. Doddington, M. A. Przybycki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, 2000.

[4] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret verification testing protocol for face recognition algorithms," Tech. Rep. NISTIR 6281, National institute of standards and technology, 1998.

[5] K. Jonsson, J. Matas, Y. P. Li, and J. Kittle, "Learning support vectors for face verification and recognition," in *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 208–213.

[6] C. Sanderson, *Automatic person verification using speech and face information*, Ph.D. thesis, School of microelectronic engineering, Griffith University, 2002.

[7] A. Albiol, L. Torres, and E. J. Delp, "Video preprocessing for audiovisual indexing," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, Fl, 2002, vol. 4, pp. 3636–3639.

[8] T. Fromherz, P. Stucki, and M. Bischsel, "A survey of face recognition," Tech. Rep. 97.01, Department of Computer Science, University of Zurich, 1997.

[9] A. Lanitis, C. J. Taylor, and T.F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743–756, July 1997.

[10] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, July 1997.

[11] S. Eickler, S. Muller, and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *Image and Vision Computing*, vol. 18, no. 4, pp. 279–287, April 2000.

[12] M.A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 1991, pp. 586–591.

[13] L. Torres and J. Vilá, "Automatic face recognition for video indexing applications," *Pattern recognition*, vol. 35, no. 3, pp. 615–625, December 2001.

[14] M. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, January 2002.

[15] A. Albiol, *Video indexing using multimodal information*, Ph.D. thesis, Universidad Politecnica de Valencia, Valencia, Spain, April 2003.

[16] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Transactions of Speech and Audio Processing*, vol. 2, pp. 639–643, October 1994.

[17] R. Balchandran, V. Ramanujam, and R. Mammone, "Channel estimation and normalization by coherent spectral averaging for robust speaker verification," in *Proceedings of the 6th European Conference on Speech Communication and Technology*, Budapest, 1999, pp. 755–758.

[18] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speakers models," *Speech Communications*, vol. 17, no. 1-2, pp. 91–108, August 1995.

[19] Q. Li, "A detection approach to search-space reduction for HMM state alignment in speaker verification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 569 – 578, July 2001.

[20] D. A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using Adapted Gaussian Mixture Models," *Digital Signal Processing. A review journal*, vol. 10, no. 1-3, pp. 19–41, Jan 2000.

[21] T.K. Moon, "The Expectation-Maximization algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, November 1996.

[22] C. Taskiran, J. Chen, A. Albiol, L. Torres, C. A. Bouman, and E.J. Delp, "Vibe: A compressed video database structured for video active browsing and search," *IEEE transactions on Multimedia*, vol. 6, no. 1, pp. 103–118, Februrary 2004.

[23] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian Mixture speaker models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 3, no. 1, pp. 72–83, January 1995.

[24] G. Wei, L. Agnihotri, and N. Dimitrova, "TV program classification based on face and text processing," in *Proceedings of the International Conference on Multimedia and Expo*, New York, USA, July 2000, pp. 1345–1348.

[25] G. Iyengar, H. Nock, C. Neti, and M. Franz, "Semantic indexing of multimedia using audio, text and visual cues," in *Proceedings of the International Conference on Multimedia and Expo*, Lausanne, Switzerland, August 2002, vol. 2, pp. 369–372.

[26] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for contend-based video segmentation," in *IEEE International Conference on Image Processing*, Chicago, IL, October 4-7 1998, vol. 3, pp. 526–530.

[27] T. J. Wark, *Multimodal speech processing for automatic speaker recognition*, Ph.D. thesis, School of Electrical & Electronic Systems Engineering, Queensland University of technology, Brisbane, 2000.

[28] P. Yin, I. Essa, and J.M. Rehg, "Boosted audio-visual HMM for speech reading," in *Proceedings of the IEEE Int. Workshop on analysis and modeling of faces and gestures*, Nice, France, October 2003, pp. 68–73.

[29] J. Luettin, *Visual speech and speaker recognition*, Ph.D. thesis, Deparment of Computer Science, University of Sheffield, 1997.

[30] P. Verlinde, "A contribution to multi-modal identity verification using decision fusion," M.S. thesis, Deparment of signal and image processing, Telecom Paris, 1999.

[31] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in *Proceedings of the 3rd International Conference on Audio- and Video-based biometric person authentication*, Halmstad, Sweden, 2001, pp. 348–353.

[32] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentlad, "Multimodal person recognition using unconstrained audio and video," in *Proceedings of the International Joint Conference on Neural Networks*, 1999.

[33] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for video broadcast news," in *Proceedings of the IEEE Signal Processing Society 1999 Workshop on Multimedia Signal Processing*, September 1999, pp. 161–167.

[34] M. Viswanathan, H.S. M. Beigi, and F. Maali, "Information access using speech, speaker and face recognition," in *Proceedings of the International Conference on Multimedia and Expo*, New York, August 2000, vol. 1, pp. 493–496.

[35] R. Brunelli, D. Favaligna, T. Poggio, and L. Stringa, "Automatic person recognition using acoustic and geometric features," *Machine vision & applications*, vol. 8, no. 5, pp. 317–325, 1995.

[36] R. D. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Willey-interscience, 2nd ed. edition, 2001.

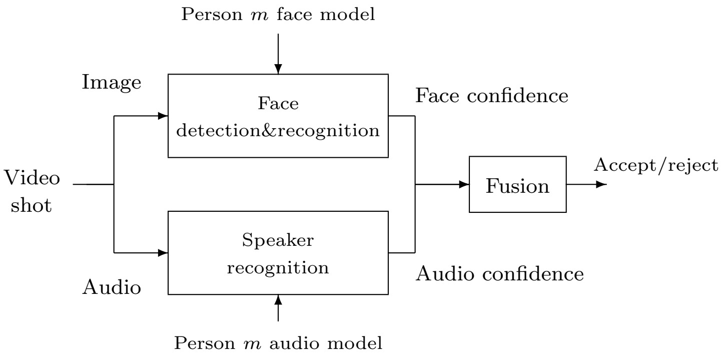


Figure 1: System overview

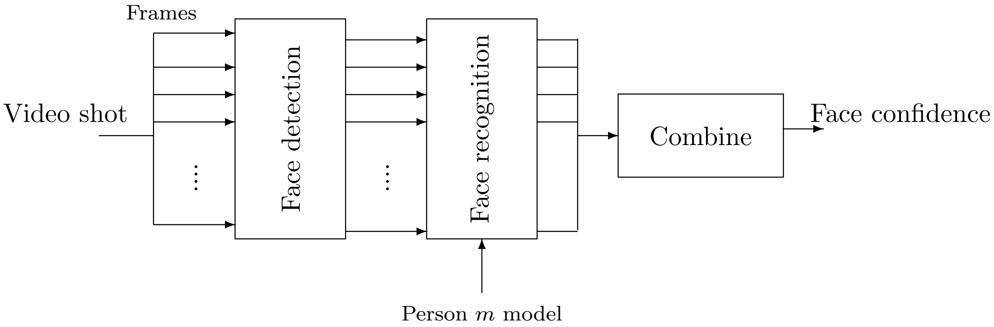


Figure 2: General diagram to extract the face confidence.

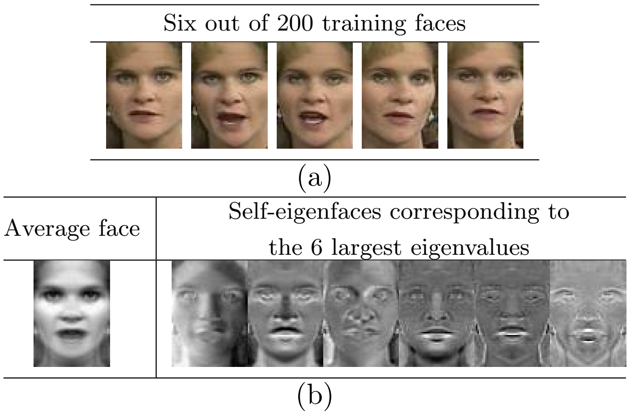

Six out of 200 training faces
(a)
Average face
Self-eigenfaces corresponding to the 6 largest eigenvalues
(b)

Figure 3: (a) Six out of 200 original training faces. (b)Average face and the first 6 self-eigenfaces obtained with the 200 training faces.

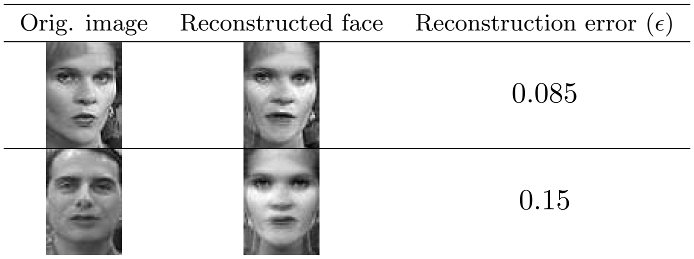


Figure 4: Example of face reconstruction using the 26 first self-eigenfaces of the person of Figure 3.a

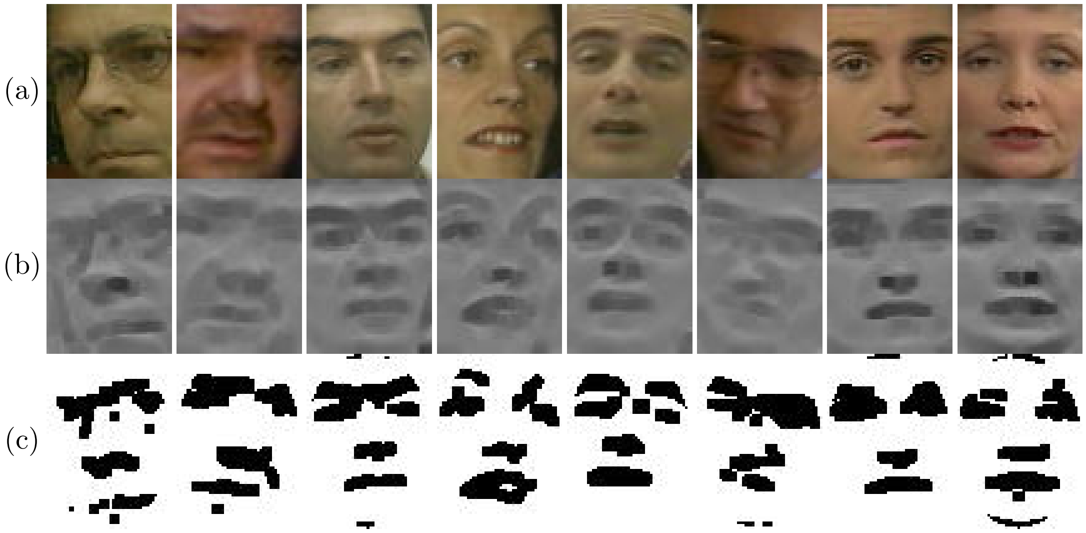


Figure 5: (a) Sample faces, (b) result of intensity normalization, (c) facial features after intensity thresholding.

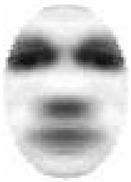


Figure 6: Result after averaging 280 face patterns as in Figure 5.c.

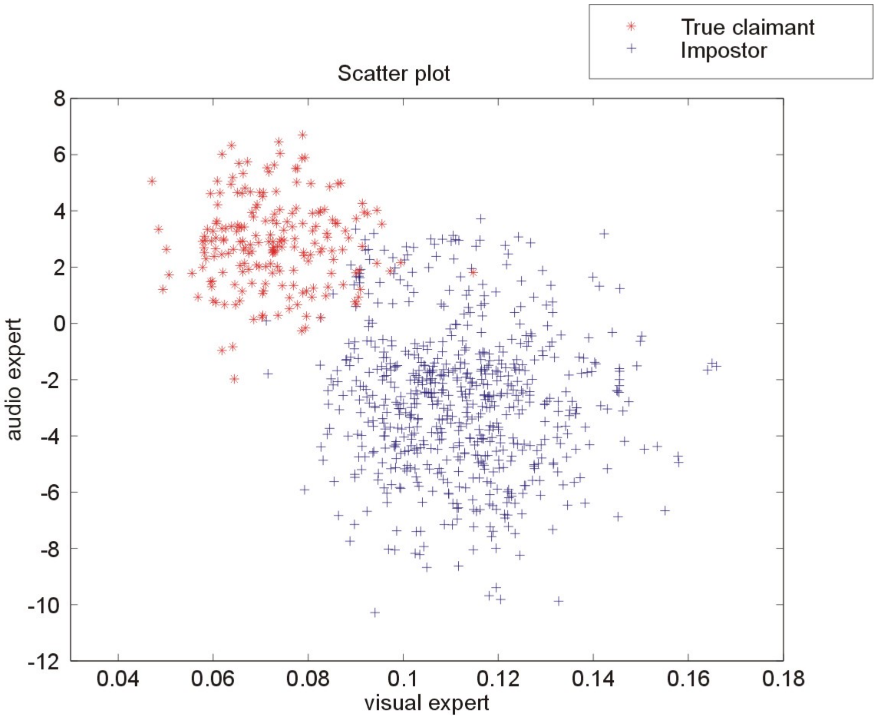


Figure 7: Scatter plot of the face and voice confidences.

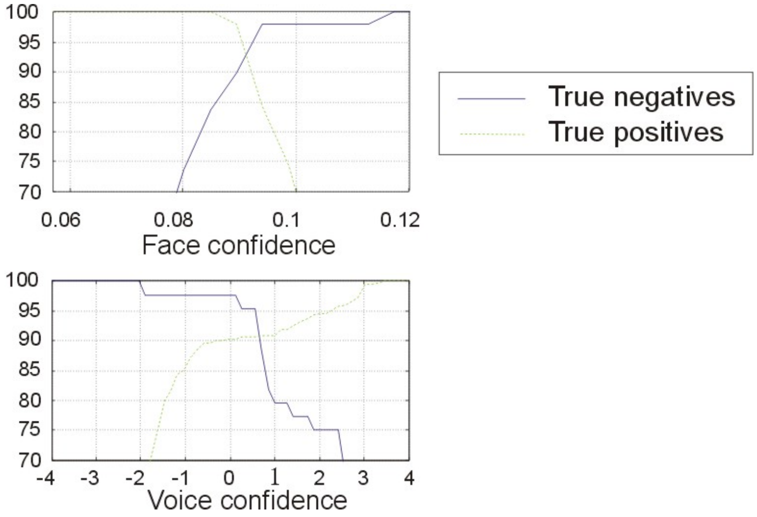


Figure 8: Recognition results using each modality separately.

	Retained variance		
	85%	90%	95 %
ID	K_m		
1	16	26	50
2	18	27	45
3	16	23	39
4	15	23	42
5	17	31	58
6	15	24	43
7	20	30	55
8	11	20	37
9	10	17	33
10	8	17	29

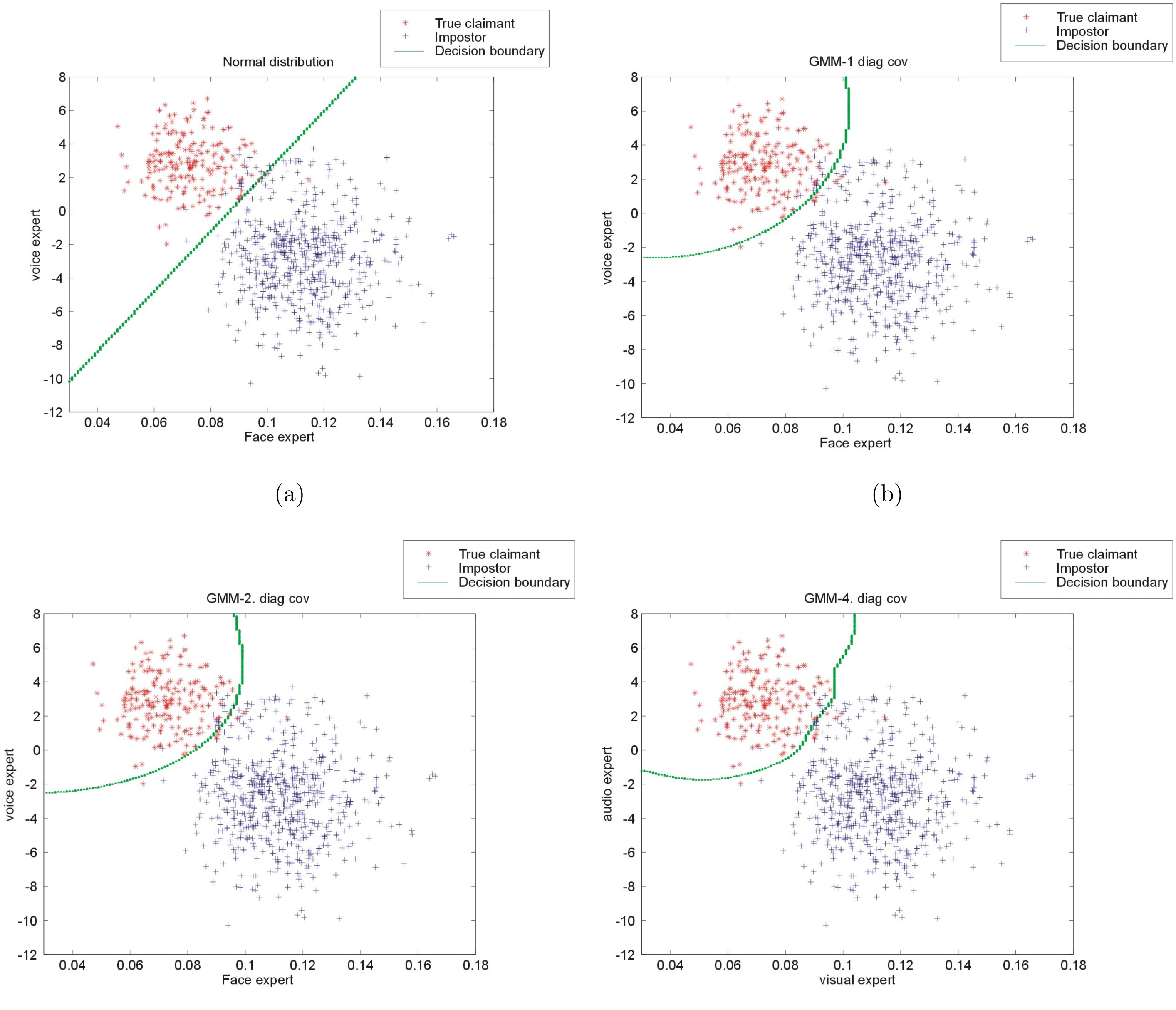
Table 1: This table shows the number of selected self-eigenfaces (K_m) for different values of retained variance for each modeled person.

Face recognition			Speaker recognition		
retained variance	threshold on FC_m	ECR	Num. of Gauss. Mix	threshold on SC_m	ECR
85 %	0.097	92.2 %	16	0.37	88%
90 %	0.091	92.8 %	32	0.56	92%
95 %	0.081	91.5 %	64	0.87	86%

Table 2: Threshold and detection rates using only the face modality

Type of cov. matrix	Model	ECR	Threshold
Diagonal	GMM-1	97.3 %	-0.43
Diagonal	GMM-2	97 %	-0.51
Diagonal	GMM-4	96.8 %	-0.13
Diagonal and equal for both classes	Normal dist.	97.1 %	0.6

Table 3: Recognition results using the Bayesian post-classifier.



(a)

(b)

(c)

(d)

Figure 9: Decision boundaries for the Bayesian post-classifier using: (a) normal distributions, (b)GMM with one mixture, (c) GMM with two mixtures (d) GMM with 4 mixt. In the three examples the covariance matrices are forced to be diagonal.

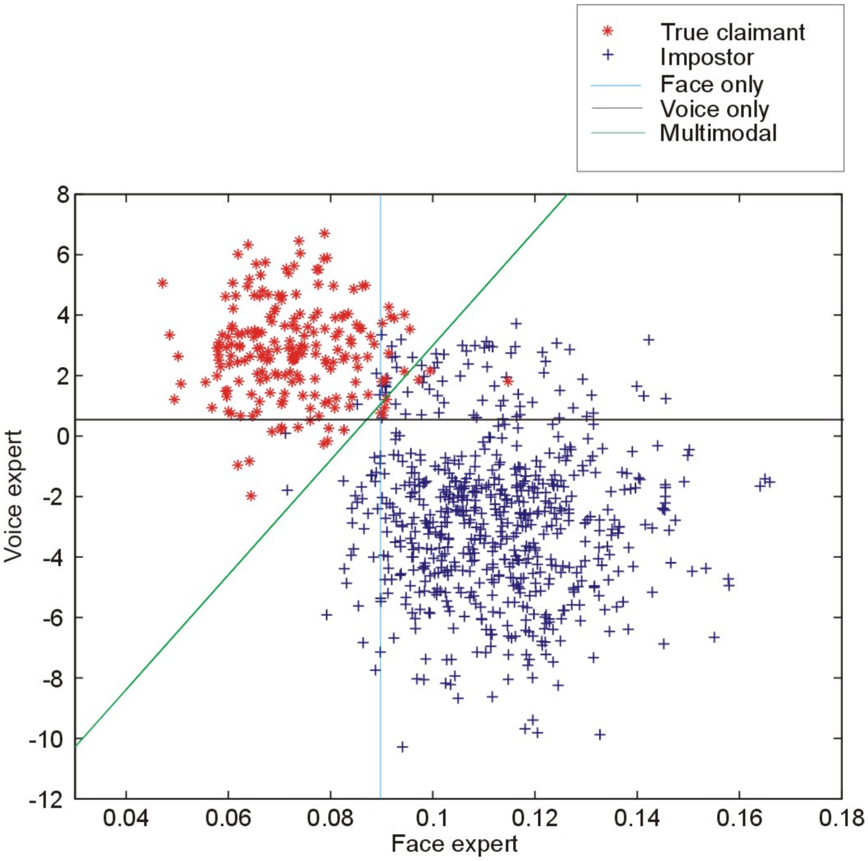


Figure 10: Comparison between decision boundaries of face only, voice only and multimodal recognition (using the classifier of Figure 9.a).