Audio to the rescue

Alberto Albiol
Technical University of Valencia, Spain
alalbiol@dcom.upv.es

Luis Torres
Technical University of Catalonia, Spain
luis@gps.tsc.upc.es

Edward J. Delp
Purdue University, USA
ace@ecn.purdue.edu

Important: Figures are located on the last page of this document

1. Introduction

Automatic recognition of people is a challenging problem which has received much attention during the recent years due to its many applications in different fields such as law enforcement, security applications or video indexing.

Recognition of people can be achieved using different biometrics such as face, voice, fingerprints or iris scans among others. In most cases, the choice of the particular biometric deeply relies on the final application. For instance, retinal scans have shown high recognition accuracy, however their use is limited to the availability of cooperative individuals, which is not always possible.

Although relatively high recognition rates have been obtained using separated biometrics, there are some limitations that make very difficult to increase the recognition performance using individual modalities. Examples of these limitations are changes in illumination or pose for the face biometric, and ambient noise and channel distortion for the voice biometric. Although, more powerful recognition schemes for each modality would probably improve the recognition rates, there is another potential way to increase the recognition performance which consists in combining recognition results from different biometrics. The key idea behind this combined approach is that different information sources can complement each other since degradations for each modality are usually uncorrelated. A good example of a system that combine multiple information sources is the human being, e.g. it has been shown that simultaneously seeing and listening a person talking greatly increases intelligibility.

In this paper, we will focus on the recognition of people in video sequences for video indexing applications. This is: given a video sequence, we want to locate those clips where a particular person *m* appears. Since we also want to use the voice information, an additional requirement is that person *m* should be also talking. Examples of this type of clips include anchors and head and shoulders sequences of people that are being interviewed in news sequences. In our approach, if person *m* is being searched, then, for each clip in the video sequence, the identity *m* will be proposed and the recognition system will verify (binary decision) this identity claim. Initially, we will discuss about person recognition using the face biometric only. Then, we will show how to improve the recognition results by using voice information. Although we focus on video indexing, the presented techniques are general enough to be applied to any other face recognition application.

The rest of this paper is organized as follows, in Section 2 a short review of face recognition techniques and details about a more specific face recognition technique suitable for video indexing are presented. Finally, an approach to combine voice and face biometrics is presented in Section 3. Finally some conclusions are drawn in Section 4.

2. Face recognition approaches

Face recognition has been an active reseach topic for more than one decade. Initially, face recognition systems focused on still images. However, during the last years research on face recognition in image sequences has gained much attention, although, nearly all systems apply still-image face recognition techniques to individual frames. The main advantage of using image sequences is that it allows to select good frames for recognition.

Face recognition approaches on still images can be broadly grouped into geometric and template matching techniques. In the first case, geometric characteristics of faces to be matched, such as distances between different facial features, are compared. This technique provides limited results although has been used extensively in the past. In the second case, face images represented as a two-dimensional array of pixel intensity values are compared to a single or several templates representing the whole face. More successful template matching approaches use Principal Components Analysis (PCA) or Linear Discriminat Analysis (LDA) to perform dimensionality reduction achieving good performance at a reasonable computational time. Other template matching methods use neural network classification and deformable templates, such as Elastic Graph Matching (EGM). Recently, a set of approaches that use different techniques to correct perspective distortion are being proposed. These techniques are sometimes referred to as *view-tolerant*. An example of these techniques is based on pseudo-2D Hidden Markow Models (HMMs) [1]. A recent comparison between different face recognition techniques can be found in a recent survey paper [2]. 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. Therefore, the technique to be used is chosen based on the final application. For instance, EGM techniques require large face resolutions. Other methods are better suited for identification applications, such as LDA, where usually only one face example is available for each person. In other cases, the difficulty of training the face model limits the use of some algorithms as in the case of the HMM algorithms. In any case, the comparison results presented in [3], over twelve different face recognition techniques used for verification applications indicate that EGM, LDA and PCA were on the top three, each method showing different levels of performance on different subsets of images.

In this paper, we are focusing on face recognition using a variant of the well known PCA technique [4] (also known as *eigenfaces*) called *self-eigenfaces* [5]. The self-eigenfaces technique is specially suitable for the video indexing applications, where, usually many training faces for the person to be recognized are available. The self-eigenface approach takes advantage of this by performing a separate PCA for each person P_i to be recognized. PCA allows a compact representation of each person P_i using the mean and the eigenvectors of the covariance matrix of the training faces of the person P_i. Since the eigenvectors of the covariance matrix look like faces, they are called *self-eigenfaces* to emphasize that they are built using different views of the person P_i. Figure 1 shows a small sample of training faces and their corresponding mean and self-eigenfaces.

Figure 1. This figure shows a small sample of training faces and the corresponding mean face and first self-eigenfaces used to model this particular person.

During the test phase, each test face is projected and reconstructed using a particular set of self-eigenfaces. Then, the reconstruction error is used as a confidence measurement that the test face

corresponds to P_i. The basic idea behind this method is that given a test face, a low reconstruction error (good fit) is achieved when the self-eigenface set of the corresponding identity is used.

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 conficence value that P_i appears in the sequence. A practical way to obtain a global confidence measurement FC(P_i), can be done using the median value:

$$\mathrm{FC}(\mathrm{P_i}) = \text{median}\,\{\mathrm{E_0(P_i)}, \mathrm{E_1(P_i)}, ..., \mathrm{E_N(P_i)}\} \quad (1)$$

where $E_k(P_i)$ is the face confidence for frame *k*. The median value assures that good recognition is obtained at least for half of the frames and also allows to deal with outliers produced by changes in pose or by missed detected faces.

As a general conclusion, 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*. This conclusion can be also extended to the general PCA approach.

3. Audio-visual recognition

As introduced in Section 1, recognition results can be further improved if different biometrics are combined. In this section, we will discuss how we can use voice information to increase the recognition performance.

Techniques for combining different information sources can be broadly grouped into *pre-mapping* and *post-mapping* fusion techniques [6]. In the first group, information is combined before any use of classifier or expert. Note that while a classifier provides a hard decision, an expert provides a confidence value on each possible decision. Pre-mapping fusion has been widely for instance in lip-reading where visual and speech features are combined to increase intelligibility. Post-mapping techniques combine information after mapping from the feature space to the opinion/decision space using either a classifier or an expert.

Pre-mapping techniques are more appropriate when the information sources are closely synchronized. However, if this is not the case they tend not to generalize well, specially if the number of features is too high. Another advantage of post-mapping fusion is that it can combine opinions from different experts, even if their outputs are not commensurate (different range values). For these reasons, we have used post-mapping fusion to combine the outputs from face and voice biometrics.

In general, person recognition techniques that use the voice as a biometric are usually referred to as *speaker recognition*. Note that the objective here is not to know *what* is being said (speech recognition) but *who* says it. Speaker recognition techniques usually formulate the problem as a basic hypothesis test, where, given a speech segment *S*, a decision wether it was spoken by person P_i has to be made [7]. The optimum test is given by the log-likelihood ratio:

$$AC(P_i) = \log\left\{\frac{p(\mathrm{S}/P_i)}{p(S/BM)}\right\} \quad (2)$$

where $p(S/P_i)$ and $p(S/BM)$ are the conditional probability density functions using the models of person P_i and background respectively, which are often modeled using Gaussian Mixture Models

(GMMs). More details about features and modeling of speakers can be found in [8]. After the speaker recognition process a confidence value $AC(P_i)$ is available which can be used together with the face confidence value $FC(P_i)$ to increase the recognition performance.

Figure 2. Scatter plot showing face and voice confidences in the likelihood space. It can be seen that true and false candidates are better classified in the two-dimensional space.

Figure 2 shows the scatter plot of the two-dimensional opinion vectors $[AC(P_i), FC(P_i)]$, where it can be clearly seen that true and false candidates are better separated in the two-dimensional space. This can be done using a post-classifier that takes the expert opinions as features in the likelihood space. The post-classifier needs not to be very sophisticated. In fact, we have found that a simple MSE linear classifier [9]. is a good compromise between accuracy and generalization.

Our experiments show that the audio-visual approach to person recognition increases the performance up to 97% of true classification, compared to 93% obtained using only the image information. Figure 3 shows two examples of false acceptance and false rejection where face recognition fails using the self-eigenfaces presented in Figure 1. In these two examples, audiovisual recognition can correctly accept or reject the identity of the test person.

Figure 3. This figure shows two examples where person recognition using only face information fails. If voice information is used the audio-visual system can correctly accept and reject both examples.

4. Conclusions

This paper has presented a review of some of the more succesful techniques applied to face recognition. It has also been shown that by including the speech information, the face recognition performance increases which proves that the combination of audio and visual information is a very promising trend in face recognition.

References

1. S. Eickler, S. Muller and G. Rigoll. Recognition of JPEG compressed face images based on statistical methods. *Image and Vision Computing,* 18(4): 279-287, April 2000.
2. W. Zhao, R. Chellappa, A. Rosenfeld and P.J. Phillips. Face recognition: A literature survey. Technical Report CART-TR-948. University of Maryland, Aug. 2002.
3. P.J. Phillips, H. Moon, S. A. Rizvi and P.J. Rauss. The FERET verification testing protocol for face recognition algorithms. Technical report NISTIR 6281, National Institute of Standards and Technology, 1998.
4. M.A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586-591, Jun. 1991.
5. E. Acosta, L. Torres and A. Albiol. An automatic face detection and recognition system for video indexing applications. In *International Conference on Acoustics, Speech and Signal Processing.* Orlando, FL, May 2002.

6. C. Neti and A. Senior. Audio-visual speaker recognition for broadcast news. In *DARPA HUB4 Workshop*, Washinton D.C., March 1999.
7. S. Furui. *Automatic speech and speaker recognition,* An overview of speaker recognition technology, pages 31-56. Kluwer Academic Publishers, 1996.
8. D. A. Reynolds, T. F. Quatieri and R. B. Dunn. Speaker verification using adapted Gaussian Mixture models. *Digital Signal Processing. A review journal*,10(1-3): 19-41, Jan 2000.
9. R. D. Duda, P.E. Hart and D.G. Stork. *Pattern Classification.* Wiley-interscience, 2nd edition, 2001.

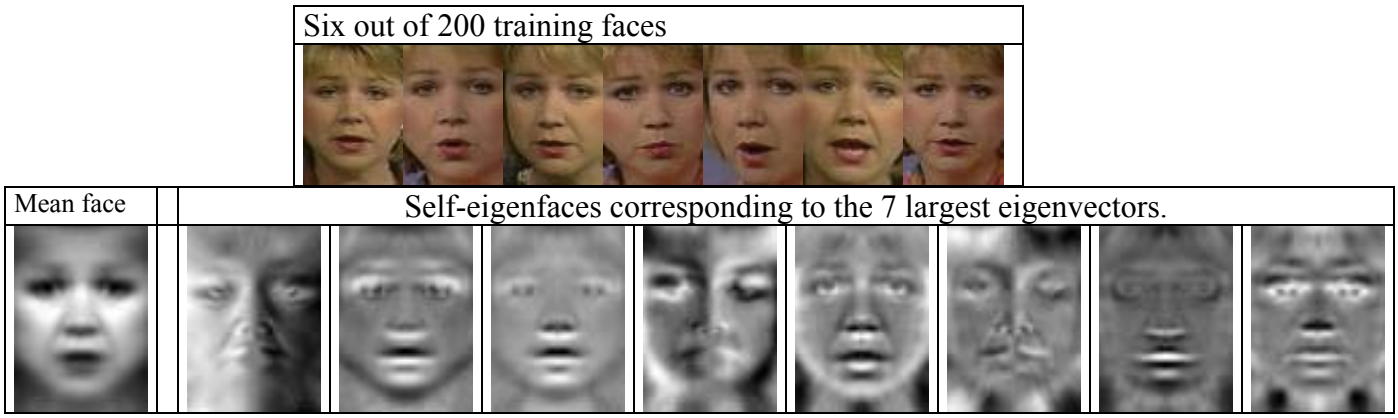


Figure 1. This figure shows a small sample of training faces and the corresponding mean face and first self-eigenfaces used to model this particular person.

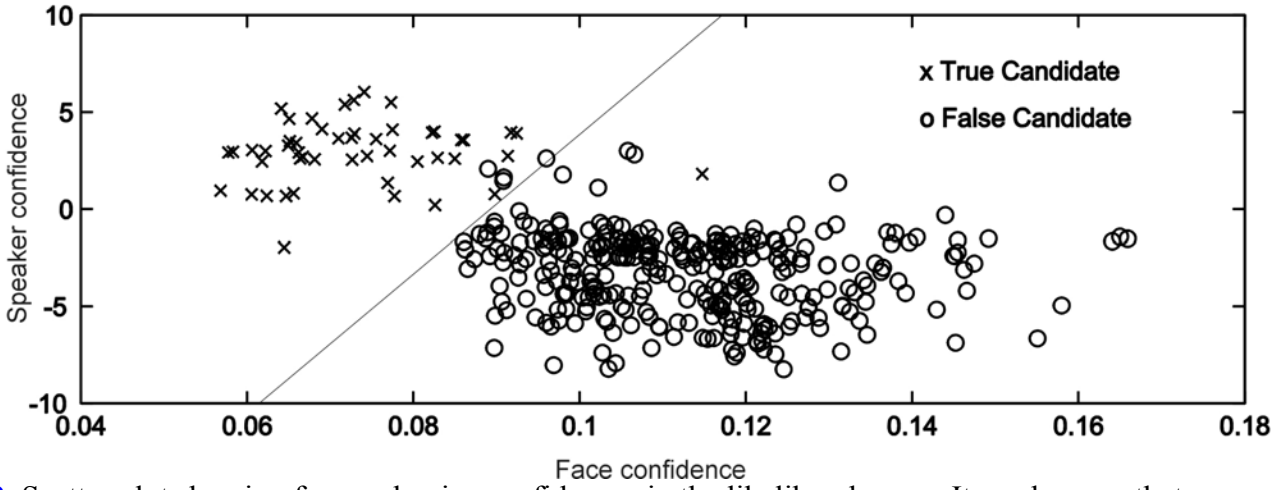
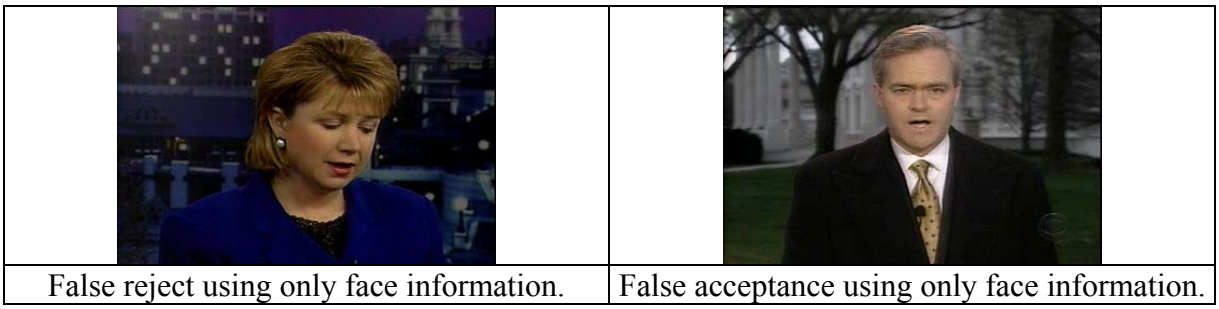


Figure 2. Scatter plot showing face and voice confidences in the likelihood space. It can be seen that true and false candidates are better classified in the two-dimensional space.



False reject using only face information.	False acceptance using only face information.

Figure 3. This figure shows two examples where person recognition using only face information fails. If voice information is used the audio-visual system can correctly accept and reject both examples.