TWO ARE BETTER THAN ONE: WHEN AUDIO COMES TO THE RESCUE OF VIDEO

Alberto Albiol

Politechnic University of Valencia, Spain
e-mail: alalbiol@dcom.upv.es

Luis Torres†*, Edward J. Delp*‡

Technical University of Catalonia, Spain†
e-mail: luis@gps.tsc.upc.es
Purdue University, USA‡
e-mail: ace@ecn.purdue.edu

ABSTRACT

This paper presents a system that automatically recognizes people in video sequences. To that end, audio and video information is used to obtain a confidence value that indicates the likelihood that a specific person appears in a video shot. Finally, a post-classifier is used 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.

1. INTRODUCTION

The needs of law enforcement and security personnel, and automated video indexing applications for video archives are driving the development of new automated systems and techniques that can automatically identify people without the assistance of human operators. These systems typically use a biometric key to 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 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.

Although relatively high recognition rates have been obtained using a single biometric, methods that use multiple biometrics can increase the system's effectiveness. A combined approach using complimentary biometrics can improve system performance because degradations for each modality usually are uncorrelated. Examples of degradations are changes in illumination or pose for a face-based biometric, or changes in ambient noise and channel distortion for a voice-based biometric.

In this paper, a system that recognizes people in video sequences combining audio and image information is proposed. More specifically we are interested in locate shots where some particular person apears in the image while talking. Examples of these shots include taped footage of news anchors, and head and shoulders sequences of people being interviewed.

In our approach, if person m is being searched, then, for each shot in the video sequence, the identity m will be proposed and the recognition system will verify or deny this identity claim. This idea is depicted in Figure 1. As shown in the Figure, audio and image information are processed in parallel and two confidence values are extracted for each selected shot. Finally, a *fusion* module is used to make the final decision based on the audio and image confidence values.

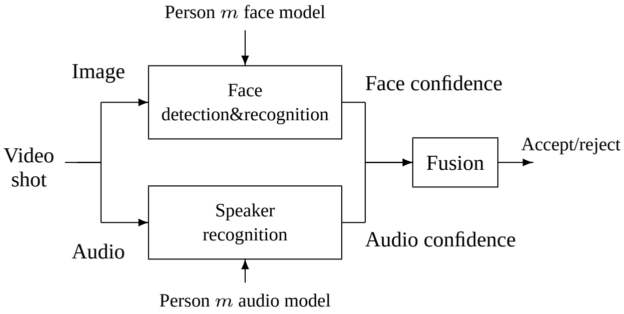


Fig. 1. System overview

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 this confidence values are used to accept or reject the identity claim. Finally, some results and conclusions are presented in Section 5.

This work has been partially supported by the grants TIC2001-0442, TIC2002-02469 of the Spanish Government and the US-Spain Joint Commission for Scientific and Technological Cooperation

2. FACE DETECTION AND RECOGNITION

2.1. Face detection

Our face detection system [1] has three main blocks. First, skin detection is used to locate regions which potentially might contain a face based on the color information. However, skin detection will likely produce non-homogeneous skin-like regions containing more than one object. The second block, unsupervised segmentation, aims to segment the skin detected pixels into a set of connected homogeneous regions containing one object. The unsupervised segmentation is performed in two stages where chrominance and luminance information are used consecutively. At each stage, we use an algorithm that combines pixel and region color segmentation techniques. However, the unsupervised segmentation can sometimes further divide the face region. Thus, region merging is used to iteratively extract a set potential face candidates. Next, we use simple constraints regarding shape, color and texture to discard many false candidates. This results in a much smaller set of face candidates for each test image, although we still might have some erroneous candidates. These candidates will finally discarded at the face recognition stage.

2.2. Face recognition

Face recognition has been an active research topic for more than one decade. Initially, face recognition systems focused on still images. In recent years, face recognition in image sequences has gained significant attention. Image sequences offer the advantage of allowing an automated system to select individual frames that offer the best chance for a biometric match with the stored video footage.

In this paper, face recognition is based on a variant of the well known PCA technique [2], also known as the self-eigenfaces technique [3]. The self-eigenfaces technique is well suited for the video indexing applications when many images of a specific face viewed from a similar perspective are available for training purposes. The main difference with the normal eigenface approach is that we model each person m with a different set of eigenfaces that we call *self-eigenfaces*. In our approach, the location of the eyes is used to normalize the size and rotation of each training view. Figure 2 presents an example the self-eigenfaces corresponding to the largest eigenvales are shown.

In the test phase, each face candidate is projected and reconstructed using a particular set of self-eigenfaces. Then, reconstruction error is used to measure the confidence that the identity of the face candidate is m. Notice that this approach is more robust to changes in brightness that the normal eigenface approach. 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. Figure 3 illustrates this idea where original and reconstructed test faces are shown. It can be seen that reconstruction error will be smaller when the identity of self-eigenfaces matches the identity of the test face.

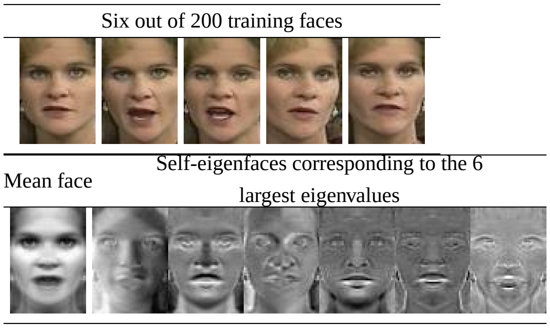

Six out of 200 training faces
Mean face
Self-eigenfaces corresponding to the 6 largest eigenvalues

Fig. 2. This figure shows and example of training faces and their corresponding self-eigenfaces.

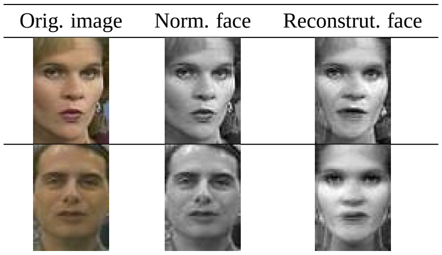


Fig. 3. Example of face recognition using the self-eigenfaces of Figure 2

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 e_i be the minimum reconstruction error for all the face candidates of frame i. Then, we define:

$$FC_m = \text{median}\{e_0, e_1, e_2, \ldots, e_N\} \tag{1}$$

as the shot-based measurement 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.

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.

3. SPEAKER RECOGNITION

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 whether or not it was spoken by person has to be made.

The optimum test is given by the log-likelihood ratio:

$$AC_m = \log\left\{\frac{p(S/P_i)}{p(S/BM)}\right\} \tag{2}$$

where $p(S/m)$ and $p(S/BM)$ are the probability density functions of the person m and the background model respectively [4]. These probability density functions are modeled using GMM's. The mixture models are built from feature vectors that consist of 12 mel-frequency cepstral coefficients and its corresponding delta coefficients. Cepstral vector coefficients are extracted each 17 ms. using a 34 ms. Hamming window. We also remove the silent frames and use cepstral mean subtraction to reduce linear channel distortions. In the training stage, we created a 32-GMM model for each person P_i using 2–3 min of clear speech. 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 [5]. To make the background model as universal as possible, we took special care on the composition of the speaker's universe, trying to balance as much as possible the number of male/female utterances and different recording conditions. In the test stage we use AC_m as the confidence measurement that the utterance was spoken by m.

4. COMBINED AUDIOVISUAL RECOGNITION

Once audio and visual confidences AC_m and FC_m are extracted, the *fusion* block of Figure 1 is used to make the final multi-modal decision.

Among many possibilites for fusing audio and visual information [6], we choose a post-classifier. The post-classifier option has several important advantages compared to other options. For instance, it is able to combine opinions from different expert classifiers, even if their outputs fall in different ranges, because it directly maps the input values from the confidence space to the decision space. Furthermore, this mapping takes into account the degree of confidence or goodness of each separate modality.

Figure 4 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.

In our experiments we used several types of classifiers such as Bayesian classifiers and MSE classifiers [7]. Our Bayesian classifiers use GMM to model the conditional densitiy functions. However, we found that the simplest classifier based on a linear discriminant function is the best option in terms of reducing the training error while achieving good generalization. The linear discriminant function is found by MSE using the Pseudoinverse Matrix [7]. Figure 4 shows the decision boundary that is found using this method.

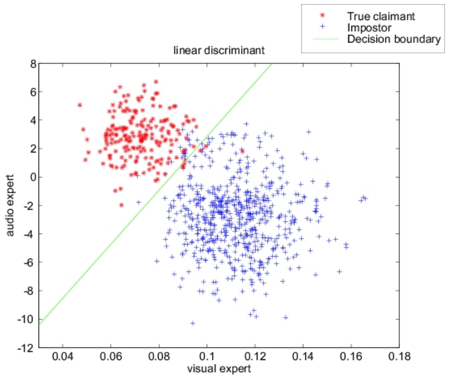


Fig. 4. Scatter plot of the face and speaker confidences.

5. RESULTS

We carried out all the experiments on several TV news sequences stored on our video data base [5].

From these sequences we selected 10 people that appear frequently. This was done for practical reasons, think that usually head and shoulders shots in TV news stories usually 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) to test if we can find the selected candidate in different sequences.

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* (following the terminology traditionally used in security access systems). If the system makes the right decision for a true claimant shot then we say that it is a *true posisite*. 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*.

It should be also mentioned that the recognition experiments were conducted only in head and shoulders shots. This was done to increase the difficulty to the recognition system, since it is relatively easier to reject shots where no faces are present or nobody is talking.

Figure 5 shows the true positives and true negatives curves for each independent modality. These results are obtained by thresholding face or speaker confidences at different threshold values. It can be appreciated that for an equal error rate (same percentage of true positives and negatives), face and speaker modalities achieve around a 93% and 91% of success respectively.

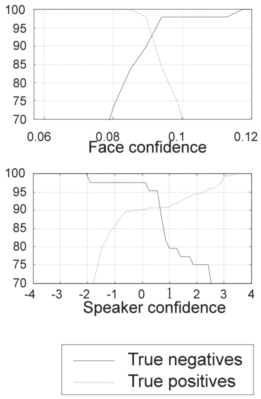


Fig. 5. Recognition results using only each modality separately.

Most of the false negatives are produced because nonfrontal views are available in a test shot, and thus a very low face confidence is obtained. On the other hand, background noise increases considerably the bias of the speaker confidence and this is probably the reason for most of the false positives and some false negatives.

The previous results can be improved if we combine the speaker and face confidences in the classifier described in section 4. Using the proposed classifier we obtain a 98% of true positives and a 99% of true negatives represents a significant improvement respect to each separate modality.

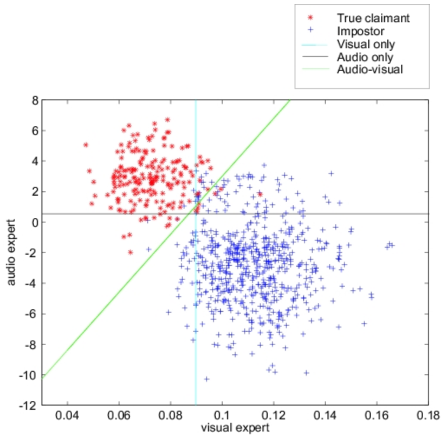
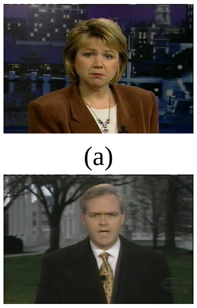


Fig. 6. Comparison between decision boundaries of visual only, audio only and audio-visual recognition.

Figure 6 shows the decision boundaries obtained by using visual only, audio only or audio-visual informations. In the Figure is clearly shown how true claimant and impostor shots are better separated when both information sources are used.

Figure 7 shows two examples where correct recognition is only achieved when audio and visual information are used together. The first case presents a false positive while in the second a false negative is presented.


(a)

(b)

Fig. 7. Correct recognition is only achieved if audiovisual information is used for a (a) true claimant shot (b) impostor shot.

This paper has presented a multimodal system for person recognition. It has also been shown that by including the speech information, the face recognition performance increases, proving that the combination of audio and visual information is a very promising trend in face recognition.

6. REFERENCES

[1] A. Albiol, L. Torres, and E. J. Delp, “A simple and efficient face detection algorithm for video database applications,” in *Proceedings of the IEEE International Conference on Image Processing*, Vancouver, Canada, September 2000, vol. 2.

[2] 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.

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

[4] D. A. Reynolds, “Speaker identification and verification using gaussian mixture speakers models,” *Speech Communications*, vol. 17, pp. 91–108, 1995.

[5] 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. (Accepted for publication), 2003.

[6] D. L. Hall and J. Llinas, *Handbook of multisensor data fusion*, CRC Press, USA, 2001.

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