VIDEO FACE RECOGNITION: A PHYSIOLOGICAL AND BEHAVIOURAL MULTIMODAL APPROACH

*Federico MATTA, Jean-Luc DUGELAY**

Eurecom Institute
2229 Route des Cretes
06904 Sophia Antipolis, France
{Federico.Matta, Jean-Luc.Dugelay}@eurecom.fr

ABSTRACT

In this article we present a multimodal system to person recognition by integrating two complementary approaches that work with video data. The first module exploits the behavioural information: it is based on statistical features computed using the displacement signals of a head; the second one is dealing with the physiological information: it is a probabilistic extension of the classic Eigenface approach. For a consistent fusion, both systems share the same probabilistic classification framework: a Gaussian Mixture Model (GMM) approximation and a Bayesian classifier. We assess the performances of the multimodal system by implementing two fusion strategies and we analyse their evolution in presence of artificial noise.

Index Terms— Identification of persons, Face recognition, Object recognition.

1. INTRODUCTION

For decades human face recognition has been an active topic in the field of object recognition. Most of algorithms have been proposed to deal with individual images, also called image-based recognition, where both the training and test sets consist of individual face images. However, with existing approaches, the performance of face recognition is affected by different kinds of variations, for example: expression, illumination and pose changes. Thus, researchers have started to look at video-based recognition, in which both training and test sets are video sequences containing the face. A detailed analysis of person recognition using still images, its performances and limitations can be found in [1, 2].

Person recognition using videos has some advantages over image-based recognition. First, the temporal information of faces can be exploited to facilitate the recognition task; for example, dynamical characteristics, which are specific to each person, like the motion of the head, the evolution of the pose or the mimic of the face. Second, more effective representations, such as 3D face models or super resolution images, can be obtained from the video sequence and used to improve the performance of the systems. Finally, video-based recognition enable learning or updating the subject model over time.

It's a common trend in literature to exploit only a part of the video information; in fact in our research experience [2, 3] and in the majority of research studies the recognition systems have been based either on the physiological information (facial appearance) either on the temporal one (facial motion). Considering the potential of these two independent modalities, the natural evolution of the video-based person recognition is directed towards the study of a multimodal recognition system, that is exploiting all the video information; consequently, we intend to investigate this original perspective and make use of the complementary nature of these two modalities, in order to develop a system with improved discriminating power and more robustness.

In this paper, we present a multimodal person recognition system, which is composed by two complementary modules. The first one [3], is based on displacement signals of a few head features, automatically extracted from the video sequence; statistical features are then computed from these signals and used for discriminating identities. The second system is a probabilistic extension of the classic Eigenface approach [4], in which the recognition task is done on a reduced face space, computed by using a Principal Component Analysis (PCA) transformation. For a consistent fusion, both systems share the same probabilistic classification framework: the characteristical head displacements and the personal variations in face space are modelled by Gaussian Mixture Models (GMM) and the classification task is solved as a Bayesian decision making problem. The experimental results show that the multimodal integration provides an important advantage in discriminating identities, expecially in presence of corrupting noise.

The main contribution of this article is the development of a multimodal framework that operates a fusion of two complementary video modalities: a well established physiolog-

*We acknowledge the SIMILAR network of excellence for funding.

ical one, related to facial appearance, and a pioneering behavioural one, related to head motion. Moreover there are some important side contibutions as: 1) the analysis on noise robustness of the two unimodal systems and the multimodal one proposed; 2) the analysis on the effectiveness of unbalanced fusion (the fusion between a performing and a weak system), which is still an open question in literature.

The rest of the paper is organized as follows: in section 2 we detail our recognition system, then we report and comment experiments in section 3, and finally we conclude this paper with remarks and future work in section 4.

2. DESCRIPTION OF THE RECOGNITION SYSTEM

Our person recognition system can be organized in three different modules: a Static Recognizer, which computes its recognition scores by using the facial appearance of the subject, a Temporal Recognizer, which computes its scores by exploiting the facial motion of the individual, and a Fusion Module, which achieves the identification and verification task by integrating the two previous modalities.

2.1. Static Recognizer module

Our appearance-based recognition algorithm is a probabilistic extension of the classic Eigenface approach, presented in [4] by Turk and Pentland.

Following the original technique, we compute the Principal Component Analysis (PCA) on a general set of face images, in order to obtain a set of orthogonal vectors (the eigenvectors) that optimally represent the distribution of the data in the Root Mean Squares (RMS) sense; these vectors define the sub-space of face images, which we call face space. A new face image is transformed into its eigenface components by a simple projection into the face space; we will denote the projected vector for image k as: $\mathbf{y}_k$.

Then, we improve the original identification and verification task by using a Bayesian framework. Firstly, for each individual we want to model the distribution of his images in the face space; we approximate the class-conditional probability density function of the individual by using a Gaussian Mixture Model (GMM). We can express using the following formula:

$$\mathrm{P}\left(\mathbf{y}_k \mid \omega_q\right) = \sum_{c=1}^{C} \alpha_c \aleph\left(\mathbf{y}_k; \mu_c, \mathbf{\Sigma}_c\right) \tag{1}$$

where ω_q refers to class (individual) q, α_c is the weight of the c-th Gaussian component, $\aleph$. After that, in order to test a given image, we compute the log-posterior probabilities for each class q (we will refer to them as β):

$$\beta_{q,k} = \log\left(\mathrm{P}\left(\omega_q \mid \mathbf{y}_k\right)\right) = \log\left(\frac{\mathrm{P}\left(\mathbf{y}_k \mid \omega_q\right)\mathrm{P}\left(\omega_q\right)}{\mathrm{P}\left(\mathbf{y}_k\right)}\right) \tag{2}$$

For completeness, we note that the priors and scaling factors, presented in the previous formula, are direclty estimated from the training database.

2.2. Temporal Recognizer module

A detailed description of this module can be found in our previous publication [3]; here we provide a short summary of the algorithm.

Head motion is firstly analysed by retrieving the displacements of the eyes, nose and mouth in each video frame. Then, the raw signals are transformed and normalized in order to obtain video independent feature vectors, in a way that the scale and geometrical parameters do not interfere with our recognition results. Finally, we model the distribution of characteristic displacements (embedded in feature vectors) over time by training individual Gaussian Mixture Models (GMM), and we achieve classification through a Bayesian classifier.

For integrating this module in our multimodal system, we recover all log-posterior probabilities (the similarity scores) computed during the Bayesian classification (we will refer to them as γ):

$$\gamma_{q,k} = \log\left(\mathrm{P}\left(\omega_q \mid \mathbf{X}_k\right)\right) \tag{3}$$

in which ω_q refers to class q and $\mathbf{X}_k$ is a matrix containing the behavioural features extracted from video k.

2.3. Fusion Module

The Fusion Module integrates the two similarity scores (log-posterior probabilities) from the previous modules and computes identification and verification rates of the multimodal system. In this paper, the multimodal similarity score is calculated by two versions of the weighted sum fusion rule [5], which in our case has the following general formula:

$$\theta_{q,k} = b_{q,k}\beta_{q,k} + g_{q,k}\gamma_{q,k} \tag{4}$$

in which $b_{q,k}$ and $g_{q,k}$ are the two weights.

In the first implementation, we compute the mean between the scores of the two modules (equal weighting):

$$b_{q,k} = g_{q,k} = 0.5 \quad \forall q, k \tag{5}$$

This simple technique has an interesting probabilistic interpretation. If we assume that $\mathbf{X}_k$ and $\mathbf{y}_k$ are independent from each other and equally distributed, and that all the classes are equiprobable (a common scenario in real applications), then the similarity score $\theta_{q,k}$ is the joint log-posterior probability of $\mathbf{X}_k$ and $\mathbf{y}_k$:

$$\theta_{q,k} = 0.5\log\left(P\left(\omega_q \mid \mathbf{y}_k, \mathbf{X}_k\right)\right) + T \tag{6}$$

where T is a constant translating factor.

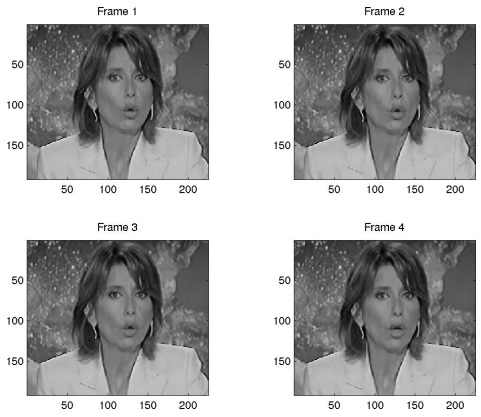

Fig. 1. The first 4 frames of a video sequence.

The second implementation of the similarity score for the integrated system is an adaptive weighting, successfully applied by Chang et al. [6], and computed as follows:

$$\begin{aligned} b_{q,k} &= \frac{\beta_k^{1st} - \beta_k^{2nd}}{\beta_k^{1st} - \beta_k^{3rd}} \\ g_{q,k} &= \frac{\gamma_k^{1st} - \gamma_k^{2nd}}{\gamma_k^{1st} - \gamma_k^{3rd}} \end{aligned} \tag{7}$$

in which β_k^i and γ_k^i are the i-th best scores for a given test k. The general idea of this weighting scheme is that, if the difference between the first and second scores is large compared to the typical one, then the modality can be considered reliable and its weight is big. In our implementation, we normalize the scores to sum to 1: $b_{q,k} + g_{q,k} = 1 \quad \forall q, k$.

3. EXPERIMENTS AND RESULTS

3.1. Video database

Unfortunately, existing standard video databases do not match the requirements for efficiently testing our algorithms: in particular, the Temporal Recognizer module needs a few minutes for each individual, in order to extract the temporal information and train the GMM models. For this reason, we collected a set of 192 video sequences of 12 different persons, for the task of training and testing our system. The video chunks are showing TV speakers, announcing the news of the day: they have been extracted from different clips during a period of 14 months. A typical sequence has a spatial resolution of 352×288 $pixels$ and a temporal resolution of 23.97 $frames/second$, and lasts almost 14 $seconds$ (refer to Figure 1 for an example). Even though the videos are of low quality, compressed at 300 $Kbits/second$ (including audio), the behavioural approach of our system is less affected by the visual errors, introduced during the compression process, than the pixel-based methods. Moreover, the videos are taken from a real case: the behaviour of the speakers is natural, without any constraint imposed to their movement, pose or action.

Method	1-Best (%)	3-Best (%)	EER (%)
SR (0)	93.75	97.92	2.37
SR (1)	90.63	97.92	3.88
SR (2)	71.88	90.63	8.66
SR (3)	55.21	81.25	15.53

Table 1. Identification and verification results for the Static Recognizer (SR).

Method	1-Best (%)	3-Best (%)	EER (%)
TR (0)	92.71	97.92	6.91
TR (0.333)	84.38	97.92	7.10
TR (1.333)	68.75	91.67	13.87
TR (3.333)	63.54	84.38	18.37

Table 2. Identification and verification results for the Temporal Recognizer (TR).

3.2. Image database

Concerning the Static Recognizer module, we created an image database derived from the video database depicted before; for each training video we extracted 28 $frames$ (2 $frames/second$), while for the testing set we retrieved only the first frame. Due to the well known high sensitivity of PCA-based recognition algorithms to facial alignment, variation in pose and scale, we manually normalized the image database by cropping the face region, then aligning and (in-plane) horizontally rotating the heads.

3.3. Experimental set-up

In our experiments, for the Temporal Recognizer module we selected 96 sequences for training (8 for each of the 12 individuals), and the remaining 96 (out of 192) were left for testing. For a detailed discussion of the parameters of this module, please refer to [3].

Concerning the Static Recognizer module, we selected a total of 2688 images for training (224 per individual) and 96 for testing; though, the algorithm is actually working with 5376 training images (448 per individual) due to vertical mirroring of original images. In our experiments we were obliged to chose a small eigenspace of dimension 13, due to the difficulty of approximating high dimensional distributions with a limited amount of training data; for the same reason we also considered GMMs with $1 \div 3$ Gaussian components for each model.

3.4. Results

The recognition results of the two disjoint modules are summarized in Tables 1 and 2, while the scores for the multimodal system using the mean and the adaptive weighting fusion operators are presented in Tables 3 and 4, respectively. The sec-

Method	1-Best (%)	3-Best (%)	EER (%)
SR (0) + TR (0)	94.79	97.92	2.18
SR (1) + TR (0.333)	92.71	97.92	2.84
SR (2) + TR (1.333)	82.29	93.75	6.77
SR (3) + TR (3.333)	68.75	86.46	12.88
SR (0) + TR (3.333)	93.75	97.91	2.32
SR (3) + TR (0)	84.38	95.83	7.58

Table 3. Identification and verification results obtained by fusioning with the mean operator.

Method	1-Best (%)	3-Best (%)	EER (%)
SR (0) + TR (0)	96.88	97.92	2.75
SR (1) + TR (0.333)	92.71	97.92	4.07
SR (2) + TR (1.333)	84.38	92.71	6.06
SR (3) + TR (3.333)	68.75	85.42	13.16
SR (0) + TR (3.333)	95.83	97.92	2.89
SR (3) + TR (0)	84.38	95.83	9.52

Table 4. Identification and verification results obtained by fusioning with the adaptive weighting operator.

ond and third columns of these tables represent the correct identification scores, when considering only the best scores and the best 3; the fourth column contains the Equal Error Rates (EER) in a verification application.

Considering only the first row of each table, it is possible to appreciate the best results of the four configurations, because there is no artificial noise corrupting them (the noise value in parenthesis is 0). From the results it is clear that the multimodal integration of the spatial and temporal systems increases the correct identification scores; we also observed a similar behaviour for the verification task.

By looking at the second, third and fourth rows of each table, it is possible to analyse the evolution of the recognition scores with the increment of artificial noise, which allows us to simulate a performance degradation in our system. For the Static Recognizer, we added a centered Gaussian noise with variable standard deviation (indicated in parenthesis with 1, 2 and 3) to each pixel; in the Temporal Recognizer case, the centered Gaussian noise is added to the displacement signals (expressed in parenthesis too). What is important to notice is that the more the performance of the unimodal systems degenerate (Tables 1 and 2), the higher is the gain after fusion (Tables 3 and 4), revealing more discriminating and robust systems; on the other hand, the two fusion methods performs very closely.

Finally, the last two rows of Table 3 and 4 represent the recognition scores for unbalanced multimodal systems, which are obtained by fusioning a performing one with a weak one. In the tables we report the results for extreme cases, those with the highest unbalance, in order to evaluate the degradation in the worst cases. From this experiments we can see that, even if the best multimodal systems are obtained by fusioning the best unimodal ones, the unbalanced fusion is not excessively degradating the scores of the best system; thus, our results support the argument that it is always better to fuse complementary systems, even if the have different performances.

4. CONCLUSION AND FUTURE WORKS

In this article, we analysed the effects of using at the same time the physiological and behavioural information, which is embedded in video data, for recognition purposes. Our experimental results show that this integration provides an important advantage in discriminating identities, and that in practical cases (those with balanced fusion) it is advantageous to fuse unimodal systems, expecially in presence of corrupting noise. However, we are aware that our work needs a bigger experimental validation and that should be extended to diverse video databases.

Our system can be improved in multiple ways. One way could be to modify the Static Recognizer module, by replacing the PCA-based approach with a more performing recognition algorithm. Then, the Temporal Recognizer module might be improved by adding facial mimic information: it could integrate the eye blinking or the lip motion with the head displacements. Finally, there is a variety of fusion techniques which can be investigated and possibly applied to our multimodal approach.

5. REFERENCES

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

[2] J-L. Dugelay, J-C. Junqua, C. Kotropoulos, R. Kuhn, F. Perronnin, and I. Pitas, "Recent advances in biometric person authentication," *27th IEEE Int. Conf. on Ac., Sp. and Audio Proc. (ICASSP2002)*, May 2002.

[3] F. Matta and J-L. Dugelay, "Person recognition using human head motion information," *4th Int. Conf. on Art. Mot. and Def.e Obj. (AMDO2006)*, vol. LNCS 4069, pp. 326–335, July 2006.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Jour. of Cogn. Neur.*, vol. 3, no. 1, pp. 71–86, 1991.

[5] C. Sanderson and K.K. Paliwal, "Identity verification using speech and face information," *Dig. Sig. Proc.*, vol. 14, no. 5, pp. 449–480, September 2004.

[6] K.I. Chang, K.W. Bowyer, and P.J. Flynn, "An evaluation of multimodal 2d+3d face biometrics," *IEEE Trans. on Patt. An. and Mach. Int.*, vol. 27, no. 4, pp. 619–624, April 2005.