Person Recognition based on Head and Mouth Dynamics

Usman SAEED
Eurecom Institute,
2229 route des Cretes, B.P. 193
06904 Sophia Antipolis, FRANCE,
Email: Usman.Saeed@eurecom.fr

Federico MATTA
Eurecom Institute,
2229 route des Cretes, B.P. 193
06904 Sophia Antipolis, FRANCE,
Email: Federico.Matta@eurecom.fr

Jean-Luc DUGELAY
Eurecom Institute,
2229 route des Cretes, B.P. 193
06904 Sophia Antipolis, FRANCE,
Email: Jean-Luc.Dugelay@eurecom.fr

***Abstract*— Face is considered as an attractive biometric but because of multiple sources of variabilities, the associated recognition rate is not high enough, when working on appearance only, for most of real applications. Considering that most available visual data are videos and not still images, we investigate in this article the possible contribution of some dynamic parameters (head displacements and mouth motion) in person recognition. Some preliminary results tend to validate this original proposal that opens some new perspectives in the possible design of future hybrid and efficient system combining appearance and dynamics of faces.**

I. Introduction

Still images have been the dominant modality for research in face recognition, but with the rapid increase in the use of video surveillance equipment and webcams, recognizing people using video sequences has started to attract the attention of the research community. As compared with still image face recognition, video person recognition offers both, new challenges and new opportunities; in fact, image sequences not only provide abundant data for pixel-based techniques, but also record the temporal information and evolution of the individual.

The field of automatic face recognition has been dominated by systems using appearance information and have completely ignored the behavioral information that can be used for discriminating identities. Then, most of these strategies have been developed using perfectly normalized image databases, but for actual applications it would be better to work on real data; for example, low quality compressed sequences or video surveillance shots.

In this paper, we propose a new person recognition system based on temporal signals of features from rigid head displacement and non rigid mouth motion. Head movement is analyzed by retrieving the displacements of the eyes, nose and mouth in each video frame. The local mouth dynamics are extracted by detecting the motion of lips as the person speaks. Statistical features are then computed from these signals, in order to characterise the motion information from the video, and used for discriminating identities; the classification task is done using a Gaussian Mixture Model (GMM) approximation and Bayesian classifier.

The rest of the paper is organized as follows: we briefly cite the most relevant works in section II, then we detail our recognition system in section III, after that we report and comment our preliminary experiments in section IV and finally we conclude this paper with remarks and future works in section V.

II. Related works

This research area combines various algorithms for tracking, detection and recognition which have been studied for quite some time but separately. For human face tracking, many different techniques have been developed, such as subspace-based methods, pixel-based tracking algorithms, contour-based tracking algorithms, and global statistics of color histograms. Likewise, there is a rich literature on face recognition published in the last 15 years [1], [2]; however, most of these works deal mainly with still images. Moreover, a great part of the video face recognition techniques are straightforward generalizations of image face recognition algorithms: in these systems, the still image recognition strategy is applied independently for each frame, without taking into account the temporal information enclosed in the sequence. Among the few attempts aiming to address the problem of video person recognition in a more systematic and unified manner, the methods by Li & Chellappa [3], Zhou *et al.* [4] and Lee *et al.* [5] are the most relevant: all of them develop a tracking and recognition method using a unified probabilistic framework. Lip detection and segmentation techniques have mostly been studied for audio-visual speech recognition [6], facial expression recognition [7] and lip-reading but very rarely for biometrics, where they have been classified as appearance based [8] or geometric feature based [9]. Our work is also somewhat related to the visual analysis of human motion, in particular with the automatic gait recognition (field of research).

III. Proposed method

Our person recognition system is mainly composed of three parts: a head feature extractor, a mouth feature extractor and a person classifier. The first and second modules analyse the input video and extract parameters from head and mouth motion, which are aligned in combined vectors; then, the person classifier module recognises identities.

A. Head feature extractor

The head feature extractor takes as an input a video shot, representing few seconds of a speaker.

The head detection part is done semi-automatically: the user must manually click on the (face) features of interest in the first frame, then a tracking algorithm continues until the end of the sequence. In fact, the displacement signals are automatically retrieved in the image plane, using a template matching technique in the RGB color space.

After that, the system applies some global transformations to the displacement signals, that are likely to normalize them and provide a better representation for the classification task. In the end, the module returns a feature vector, computed using the horizontal and vertical displacements; the reader can find further details of the algorithm in [10].

B. Mouth feature extractor

As the requirement for our case was to validate whether mouth dynamics can play a role in person recognition, we have exploited the rough localisation of the mouth provided by the head feature extractor and developed a simple algorithm based on a combination of image processing techniques to detect the outer lip contour and then extracting geometric feature which can be useful for person recognition. Several color transforms have already been proposed for lip enhancements. As we did not expect to use only color information, we selected the color transform proposed by [11] based on the principle that blue component has reduced role in lip / skin color discrimination. It has been defined as:

$$I = \frac{2G - R - 0.5B}{4}$$

Next, working in this transformed color space, the lip outer contour is detected by using Sobel edge detector and Otsu's thresholding. As we are primarily working on a window based on the location of mouth provided by the tracking algorithm which may include other features such as the nose tip, several additional steps have to be carried out which include, dilating the image and filling in the holes, removing 8-connected components connected to the window's boundary to remove these objects and improve the shape of the lip detected by edge map.

Once the outer lip contour has been detected (refer to Figure 1) several geometric features are extracted. Lip detection being an intricate problem is prone to errors, especially the lower lip edge as reported by [12]. We faced two types of errors and propose appropriate error recovery techniques. The first type of error, which was commonly observed, was caused when the lip was missed altogether and some other feature was selected, this error can easily be detected and corrected by applying feature value and locality constraints such as the lip cannot be connected to the window's boundary and cannot have an area value less than one-third of the average area value in the entire video sequence. The second type occurs when the lip is not detected in its entirety, e.g. missing the lower lip, such errors are difficult to detect and can only be partially corrected by a temporal smoothing filter. Finally the extracted features are arranged in a mouth feature vector and used for recognition in the next step.

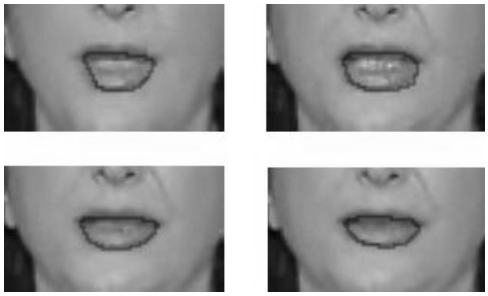

Fig. 1. Extracted lip countour.

C. Person recogniser module

The last module exploits the individual feature vectors extracted from video sequences for classification purposes.

The processed head displacements and geometric features of the mouth are firstly merged in extended global feature vectors, which are subsequently used for training a Gaussian Mixture Model (GMM) for each person in the database, in order to estimate the class-conditional probability density functions in a Bayesian classifier. More formally, if $\mathbf{x}_k$ represents the combined feature vector, then the posterior probability for class ω_q is calculated using:

$$\mathrm{P}\left(\omega_q \mid \mathbf{x}_k\right) = \frac{\mathrm{P}\left(\mathbf{x}_k \mid \omega_q\right)\mathrm{P}\left(\omega_q\right)}{\mathrm{P}\left(\mathbf{x}_k\right)}$$

The priors and scaling factors are estimated from the training set; in our experiments we have the same amount of videos for each individual so they are constant and not affecting the posterior probability computation. The class-conditional probability functions of each frame, $\mathrm{P}\left(\omega_q \mid \mathbf{x}_k\right)$, are approximated using a Gaussian Mixture Model (GMM); in formulas:

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

where α_c is the weight of the c-th Gaussian component, $\aleph\left(\mathbf{x}_k; \mu_c, \mathbf{\Sigma}_c\right)$. These class-conditional probabilities represent the likelihood that a given person has a particular head displacement and mouth opening in a given frame.

The global video score is then defined by the overall posterior probability, which is computed from the frame probabilities by making the assumption that displacements are independent (which is actually not true for our case) and by taking the product of individual probabilities,

$$\mathrm{P}\left(\omega_q \mid \mathbf{X}\right) \simeq \prod_{k=1}^{K} \mathrm{P}\left(\omega_q \mid \mathbf{x}_k\right)$$

in which $\mathbf{X}$ represents the global video feature vector.

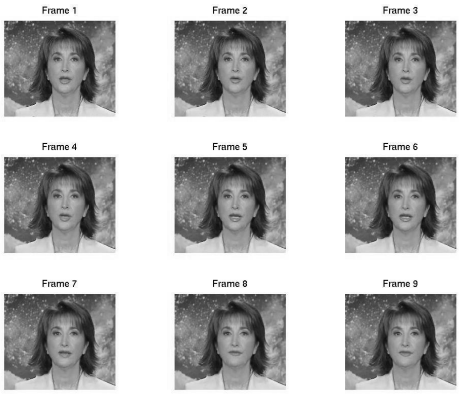

Fig. 2. The first 9 frames of a video sequence.

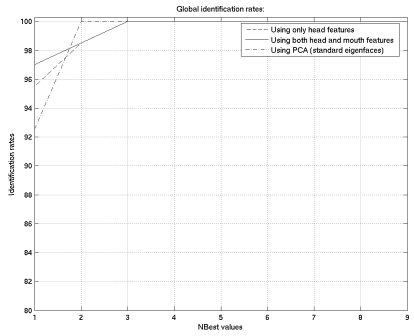

Fig. 3. Identification rates as a function of NBest values; for computing the scores, an individual is correctly identified if it is within the NBest matches.

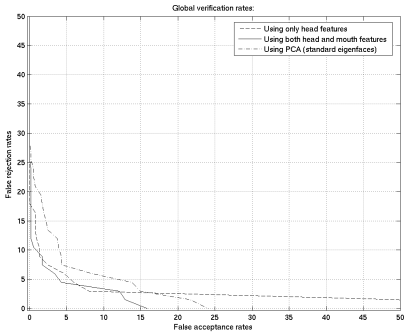

Fig. 4. Verification scores: False Rejection Rates (FRR) plotted as a function of False Acceptance Rates (FAR).

IV. Experimental results and discussion

A. Data collection

Due to the lack of any standard video database for evaluating video person recognition algorithms, we collected a set of 130 video sequences of 9 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 9 months. A typical sequence has a spatial resolution of $352 \times 288\ pixels$ and a temporal resolution of $23.97\ frames/second$, and lasts almost $14\ seconds$ (refer to Figure 2 for an example). Even though the videos are low quality, compressed at $300\ Kbits/second$ (including audio), 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.

B. Experimental set-up

For our experiments, we selected 63 video sequences for training (7 for each of the 9 individuals), and the remaining 67 (out of 130) were left for testing. From the head dynamic module we extract 8 parameters using the (horizontal and vertical) displacements of the following head features: the eyes, nose and mouth. For the mouth 4 geometric features are extracted, which include the area, the major and minor axis of the convex hull characterizing the lips and its eccentricity. In the end, we obtain a combined vector of dimension 12.

For the tracking process, keeping the initial template has showed the best discriminating properties, even if the process is not always returning the correct match (due to the absence of update); knowing the computational burden of a full template matching, we optimized the search window by taking into account the previous matches and consequently analyzing only small regions of the video frame ($74 \times 74\ pixels$).

Concerning the implementation of the head module, the displacement signals are centered around their gravity mass; more practical issues are discussed in [10]. It is important to notice that in our case all the videos have almost equal head sizes and zooms, so there is no need for spatial scaling.

Regarding the mouth dynamics and keeping normalization in mind, eccentricity was calculated from the major and minor axis of the mouth and used as an additional feature.

For training the individual GMMs, we obtained the best results using a classical Expectation-Maximization (EM) algorithm and considering 4 Gaussian components for each model. In our experiments, we were not able to add more than 9 components, because our small video database was insufficient for a reliable training of the GMMs; moreover, more complicated algorithms, which are automatically selecting the optimal number of components like the Figueiredo-Jain or the Greedy-EM [13], did not provide any advantage over the standard EM.

C. Identification and verification results

Figure 3 shows the identification scores of our system: the identification rate is 97.0%, when considering the best match ($NBest = 1$), and 100.0% when considering the three best matches ($NBest = 3$). Figure 4 shows the Receiver Operating Characteristic (ROC) curve of our system, with False Rejection Rates (FRR) plotted as a function of False Acceptance Rates (FAR): the Equal Error Rate (EER) value is 4.4%.

For providing a general reference to our experiments, we tested our video database using a pixel-based recognition

system that implements a classic eigenface algorithm. The face database for the enrollment was built from the respective video database, by extracting 13 keyframes from each video chunk; on the other hand, only one keyframe was used in the testing phase. It is important to underline that the original keyframes have been manually normalized, by cropping the faces, then aligning and horizontally warping the heads. The results have been obtained considering an eigenspace of dimension 25 and some light preprocessing. The identification rate for the best match is 92.5%, rising up to 100.0% when considering the best three matches; the equal error rate of the system is 7.2%.

Another relatively better but not an exact comparison can be observed with the work done in [10], using only head dynamics. Looking at Figure 3 and Figure 4, there is a visible improvement of about 3% on identification rates and a small improvement also on verification scores, but the significance of these improvements has to be verified by using some confidence measures, like the McNemar's tests. Another advantage of our system is that it has been applied in real cases, with compressed video sequences and no constraints on movements or actions; our behavioral approach also showed a great tolerance to face changes, due to presence of glasses and beard, or difference in haircuts, illumination and skin color. On the other hand, our technique is sensitive to within-subject variations: individuals may change their characteristic head motion when placed in different contexts or affected by particular emotional states.

V. Conclusions and future work

This pioneering work on person recognition using head and mouth dynamics showed that both global and local face dynamics may be useful for discriminating people. Our study represents a first step in the exploration of the head and mouth dynamics and their potential use in real recognition applications, either as an alternative to physical aspects of the face, like its appearance, or jointly with them.

Several improvements can be made to our system by researching and implementing different solutions. One major improvement could be to focus on the analysis of various other behavioral aspects of the human face such as blinking of eyes or motion of the pupils. This approach may show more important discriminating power, capturing the details of personal movement. Another possibility is to use our biometric system, based on head and mouth displacements, and integrate it in a multimodal one; for this purpose it could be possible to couple it with a physical modality such as appearance. It may be also interesting to refine the signal extraction process i.e. implementing a more robust tracking algorithm than the RGB template matching for head dynamics and making use of the huge literature available for lipreading to improve mouth dynamic features.

Although it is a rational thought that more precise signals could provide better classification power, the quality of those already extracted, given the accuracy of the database, is actually good enough. Meanwhile one must also consider that in the absence of constraints, the lack of prior information on the evolution of the motion and the relatively small size of the training database could be overwhelming the results. Finally, all our identification and verification results should be validated on a bigger database.

References

[1] R. Chellappa, C. L. Wilson and S. Sirohey, "Human and machine recognition of faces: a survey", in *Proceedings of the IEEE*, May 1995, College Park, USA, pp. 705–741.

[2] W. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, "Face Recognition: A Literature Survey", in *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[3] B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video", in *IEEE Transactions on Image Processing*, vol. 11, no. 5, pp. 530–544, May 2002.

[4] S. Zhou, V. Krueger and R. Chellappa, "Probabilistic recognition of human faces from video", in *Computer Vision and Image Understanding*, vol. 91, no. 1–2, pp. 214–245, July-August 2003.

[5] K. Lee, J. Ho, M. Yang and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds", in *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 303–331, September 2005.

[6] G. Potamianos, C. Neti, J. Luettin and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Overview", in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), MIT Press (In Press), 2004.

[7] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[8] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition", in *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[9] I. Matthews, G. Potamianos, C. Neti and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR", in *Proceedings of the International Conference on Multimedia and Expo*, 22–25 August 2001, Tokyo, Japan, pp. 825–828.

[10] F. Matta and J-L. Dugelay, "A behavioural approach to person recognition", to appear in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME2006)*, 9–12 July 2006, Toronto, Canada.

[11] U. Canzler and T. Dziurzyk, "Extraction of Non Manual Features for Videobased Sign Language Recognition", in *Proceedings of the IAPR Workshop on Machine Vision Application (MVA2002)*, 11–13 December 2002 , Nara, Japan, pp. 318–321.

[12] F. Bourel, C. C. Chibelushi and A. A. Low, "Robust Facial Feature Tracking", in *Proceedings of the 11th British Machine Vision Conference*, 10–13 September 2000, Bristol, England, vol. 1, pp. 232–241.

[13] P. Paalanen, J. K. Kmrinen, J. Ilonen and H. Klviinen, "Feature Representation and Discrimination Based on Gaussian Mixture Model Probability Densities - Practices and Algorithms", in *Research report of the Lappeenranta University of Technology*, no. 95, 1995.