PERSON RECOGNITION FORM VIDEO USING FACIAL MIMICS

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

Video based facial recognition is an appealing modality in biometrics due to its acceptability and ease of use but the associated recognition rate is not high enough due to multiple sources of variation and lack of constraints in real world applications. In this article we investigate the possible contribution of facial mimics extracted from low quality videos for person recognition. The initial results reported tend to validate this original proposal, thus opening some new perspectives for design of future hybrid and efficient system combining facial appearance and dynamics.

Index Terms — Image Processing, Face Recognition.

1. INTRODUCTION

Face recognition provides various advantages over other biometrics, such as acceptability and ease of use, but due to the current trends, the identification rates are low as compared to more traditional biometrics, such as fingerprints. The current trends in face recognition focus on using still images, 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. Videos not only provide abundant data for pixel-based techniques, but also record the temporal information.

The other notable trend in the field of automatic face recognition has been the use of only appearance information and thus completely ignoring the behavioral information that can be used for discriminating identities. Finally most of these techniques have been developed using perfectly normalized databases, but in a real world situation such data may not be available, thus 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 global and non rigid local motion. The local face dynamics consist of features relating to non rigid eyes and mouth motion. Rigid face movement is analyzed by calculating the degree of face symmetry and angle of the face in each video frame. Statistical features are then computed from these signals, in order to characterize 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 2, then we detail our recognition system in section 3, after that we report and comment our results in section 4 and finally we conclude this paper with remarks and future works in section 5.

2. PREVIOUS WORK

Person recognition systems that exploit temporal information in videos are recently emerging to the research community; due to the fact that these algorithms involves a heterogeneous mixture of techniques, it is challenging to classify these systems based purely on what types of techniques they use for feature extraction or classification. We therefore propose the following categories.

Holistic approach: This family of techniques analyse the head as a whole, by extracting the head displacements or the pose evolution. In [1] Li *et al*. propose a model-based approach for dynamic object verification and identification using videos. In 2002, Li and Chellappa [2] were the first to develop a generic approach for simultaneous object tracking and verification in video data, using posterior probability density estimation through sequential Monte Carlo methods. Huang and Trivedi in [3] describe a multi-camera system for intelligent rooms, combining PCA based subspace feature analysis with Hidden Markov Models (HMM).

Liu and Cheng [4] propose a recognition system based on adaptive HMMs. They first compute low-dimensional feature vectors from the individual video frames by applying a Principal Component Analysis (PCA); next they model the statistics of the sequences and the temporal dynamics using a HMM for each subject.

In [5] Aggarwal *et al.* have modelled the moving face as a linear dynamical system using an autoregressive and

moving average (ARMA) model. The parameters of the ARMA model are estimated for the entire database using the closed form solution. Recently, Lee *et al.* [6] developed a unified framework for tracking and recognition, based on the concept of appearance manifold. In this approach, the tracking and recognition components are tightly coupled: they share the same appearance model.

Feature based approach: The second group of methods exploit the individual facial features, like the eyes, nose, mouth and eyebrows. One of the first attempts to exploit facial motion for identifying people is presented by Chen *et al.* in [7]. In their work, they propose to use the optical flow extracted from the motion of the face for creating a feature vector used for identification.

Hybrid approach: These techniques use both holistic and local features. Colmenarez et al. in [8] have proposed a Bayesian framework which combines face recognition and facial expression recognition to improve results; it finds the face model and expression that maximizes the likelihood of the test image.

3. PROPOSED METHOD

Our person recognition system is mainly composed of three parts: a global feature extractor, a local feature extractor and a person classifier. The first and second modules take as input tracking points described in [9] and video sequences, to extract feature parameters. These parameters are then combined and used in person classifier module for recognition of identities.

3.1. Global feature extractor

This module is responsible for extraction of parameters relating to the rigid head motion. The parameters extracted are as under.

3.1.1. Face Angle

Face angle is defined as the angle the face makes with the horizontal axis of the image plane. Using the tracking points, the slope between the nose point $P_n(x_n,y_n)$ and the mouth point $P_m(x_m,y_m)$ is calculated as:

$$Slope(m) = \frac{(y_m - y_n)}{x_m - x_n}$$

The angle of the face with the horizontal image axis is calculated by taking the inverse tangent of the slope.

3.1.2. Face Symmetry

The second parameter extracted is the face symmetry, defined as the degree of difference between the left and right

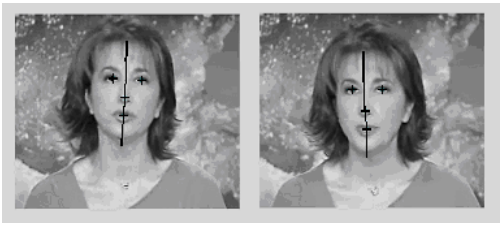

Figure 1: Facial feature points with face angle.

sides of the face. To remove the background we used an approximation of an ellipse, taking the nose point as its center. Using the nose $P_n(x_n,y_n)$, mouth $P_m(x_m,y_m)$ and eye points $Pe_1(xe_1,ye_1)$, $Pe_2(xe_2,ye_2)$ the major and minor axis are defined as:

$$MajorAxis(Maj) = \left(y_m + \frac{(y_m - y_n)}{2} \right) - \left(ye_1 + \frac{(y_m - y_n)}{2} \right)$$

$$MinorAxis(Min) = \left(xe_2 + \frac{(xe_2 - xe_1)}{8} \right) - \left(xe_1 + \frac{(xe_2 - xe_1)}{8} \right)$$

All image pixels outside the ellipse as defined above are set to zero and then from the slope, the equation of line is calculated as:

$$y - y_m = m(x - x_m)$$

This line is used as the boundary of the left and right side of the face. Depending on the fact the slope is positive or negative, one side of the face is selected. The selected side is flipped and the aligned with other side. The alignment consists of rigid rotation and translation. Once both sides are aligned the normalized MSE between the two corresponding sides is calculated.

3.2. Local feature extractor

This module extracts the local non-rigid features relating to the motion of the mouth and eyes.

3.2.1. Mouth Dynamics

In this module we have exploited the rough localization of the mouth provided by the tracking points and developed a simple algorithm based on a combination of image processing techniques to detect the outer lip contour. Several color transforms have already been proposed as in [10] for lip enhancements, we build our system on a color transform 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 working on a window

provided by the tracking algorithm which may include other objects, 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 boundary of window. Once the outer lip contour has been detected (refer to Figure 2) several geometric features are extracted, which include length of major and minor axis of lip, eccentricity and area contained within the lip contour.

We also faced two types of errors and propose appropriate error recovery techniques. The first type of error, which was observed more commonly, 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. 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.

Figure 2: Extracted lip contour.

3.2.2 Eye Dynamics

Eye movement coupled with eye brow motion are often used by humans as a means of expression to augment verbal communication, thus after the motion of the mouth, eyes provide the most valuable amount of motion information. This module extracts parameters related to the motion of the eye which includes estimation of pupil motion and blinking.

The blinking algorithm is based on a combination of image subtraction and optical flow calculation. First based on the tracking points a region of interest is selected around each of the eye. Using the assumption that global change is minimal between two consecutive ROIs, consecutive frames are subtracted to detect change. A threshold equal to twice the mean error value in the entire sequence is fixed to detect significant change, but this change can also be due to other phenomena such as global change or error in tracking, thus optical flow was also calculated using the Lucas Kanade method [12].

A mean motion vector was then calculated to have an estimate of the overall motion in the ROI. Finally the entire sequence was searched for downward, stationary and upward motion, stationary stage being when the eye is closed during blink. The sequences thus selected are considered as blinks.

The next module estimates the location of the pupil for use as a parameter. The main principle of the algorithm is that the pupil is the darkest region in the eye ROI. The algorithm is defined as:

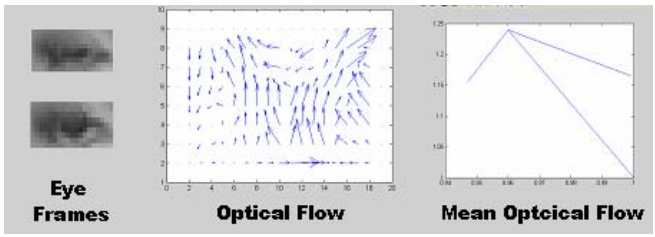

Figure 3: Optical flow of eye motion.

1. Detect the global minima in the ROI.
2. Select points with value close to global minima as possible pupil candidates.
3. Convolve candidates with a circular mask of 5 pixels.
4. Select pixel with max response as pupil center.

Circular mask of 5 pixels corresponds to the average size of pupil in our database. This parameter was not included in the experiments due to the fact that our database consists of news reporters who are generally looking at the camera or the prompter.

3.3. Person recognizer module

The last module developed by [9] exploits the individual feature vectors extracted from video sequences for classification purposes. The local and global features are firstly concatenated in various combinational vectors, which are subsequently used for training a Gaussian Mixture Model (GMM) for each person in the database. Classification is then performed by calculation the class-conditional probability density functions in a Bayesian classifier.

4. EXPERIENTAL RESULTS AND DISCUSSIONS

The database used by us for verifying our algorithms was collected by Matta and Dugelay [9], the reason for selecting this database was that previously developed databases were either not available or not suitable. This consists of 144 video sequences of 9 different TV news speakers. The videos are of low quality, compressed at 300K bits/second (including audio), which hinder in exact extraction of local facial feature.

Several combination of the feature vectors were tested to ascertain which features played a dominating role in recognition, and for providing a general reference to our experiments, we tested the video database using a classic eigenface algorithm. 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 following table describes the combination of

vectors used and the identification results obtain. The basic vector consists of the tracking points and NB refers to the number of best results used, thus NB2 means that 2 best matches were used for the identification rate calculation.

Feature Vectors	Identification Rates %		
	NB 1	NB 2	NB 3
Eigen Face algorithm	92.50	100	100
Basic Vector + Blinking	95.00	98.75	98.75
Basic Vector + Lip Vector	97.75	98.50	100
Basic Vector + Face Symmetry	91.25	97.50	100
Basic Vector + All Vectors	90.00	96.25	97.50

Table 1: Results for feature vectors in identification rates.

As it is evident from the table above that the results of almost all combination of the vectors are above 90 % identification rate. The best performance (97 % identification) is exhibited by using the lip feature vector; this was expected as the database consisted of TV news reporters who are mostly talking. The Second best performance was shown by blink vector (95 % identifications).

Using Face angle exhibited the worst results; a possible explanation could be lack of data as each person mostly moved his head by a few degrees. This problem was extended to the experiment that was carried out using all the features, which resulted in an identification rate of 90 %.

The main point 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 should be implicitly tolerant to face changes, due to presence of glasses and beard, or difference in haircuts, illumination and skin color.

5. CONCLUSIONS AND FUTURE WORKS

This nascent study explores the possibility of using facial mimics for person recognition. The preliminary results provide insight into the comparative potential of each feature, but the environment and nature of application must be kept in mind. In this study the highest identification rate was achieved when using the lip vector, which is natural as the database consisted of TV news reporters.

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 eyebrow motion. Another possibility that can surely improve results is refining the signal extraction process e.g. using snakes or deformable templates for local feature. It may be also interesting to use our biometric system, based on facial mimics and integrate it in a multimodal one; for this purpose it could be possible to couple it with a physical modality such as appearance.

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 of higher quality videos.

REFERENCES

[1] Li B., Chellappa R., Zheng Q., and Der S., "Model-based temporal object verification using video," *IEEE Transactions on in Image Processing*, Vol. 10, pp. 897-908, 2001.

[2] Li B., and Chellappa R., "A generic approach to simultaneous tracking and verification in video," *IEEE Transactions on Image Processing,* Vol. 11, pp. 530-544, 2002.

[3] Huang K. S., and Trivedi M.M., "Streaming Face Recognition using Multicamera Video Arrays," *Proceedings of 16th Inte'l Conf. on Pattern Recognition,* Vol. 4, pp. 213-216, 2002.

[4] Liu X., and Cheng T., "Video-based face recognition using adaptive hidden Markov models," *Proc. of Conf. on Computer Vision and Pattern Recognition,* Vol. 1, pp. 340-345, 2003.

[5] Aggarwal G., Chowdhury A.K.R., and Chellappa R., "A System Identification approach for Video-based Face Recognition," *Proceedings of the International Conference on Pattern Recognition,* Vol. 4, pp. 175-178, 2004.

[6] Lee K., Ho J., Yang M., and Kriegman D., "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding,* Vol. 99, pp. 303-331, 2005.

[7] Chen L., Liao H., and Lin J., "Person identification using facial motion," *Proceedings of International Conference on Image Processing,* pp. 677-680, 2001.

[8] Colmenarez A., Frey B., and Huang T.S., "A Probabilistic Framework for Embedded Face and Facial Expression Recognition," *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* Vol. 1, pp. 592-597, 2003.

[9] Matta F., and Dugelay J-L., "A behavioral approach to person recognition," *Proceedings of IEEE International Conference on Multimedia and Expo*, 2006.

[10] Canzler U., and Dziurzyk T., "Extraction of Non Manual Features for Video based Sign Language Recognition," *Proceedings of the IAPR Workshop on Machine Vision Application,* pp. 318–321, 2000.

[12] Lucas B.D., and Kanade T., "An iterative image registration technique with an application to stereo vision,"*Proceedings of Imaging understanding workshop,* pp. 121-130, 1981.