Person Recognition using Human Head Motion Information

Federico MATTA[1] and Jean-Luc DUGELAY[1]

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

Abstract. This paper describes a new approach for identity recognition using video sequences. While most image and video recognition systems discriminate identities using physical information only, our approach exploits the behavioural information of head dynamics; in particular the displacement signals of few head features directly extracted at the image plane level. Due to the lack of standard video database, identification and verification scores have been obtained using a small collection of video sequences; the results for this new approach are nevertheless promising.

1 Introduction

In the past few decades, there has been intensive research and great strides in designing and developing algorithms for face recognition from still images; only recently the problem of recognizing people using video sequences has started to attract the attention of the research community. Compared with conventional still image face recognition, video person recognition offers several new challenges and opportunities; in fact, image sequences not only provide aboundant data for pixel-based techniques, but also record the temporal information and evolution of the individual.

The area of automatic face recognition has been dominated by systems using physical information, such as greylevel values; while these systems have indeed produced very low error rates, they ignore the behavioural 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 common data; for example, low quality compressed sequences or video surveillance shots.

In this paper, we propose a new person recognition system based on displacement signals of a few head features, automatically extracted from a short video sequence. Instead of tracking the head as a whole, its movement is analysed by retrieving the displacements of the eyes, nose and mouth in each video frame. Statistical features are then computed from these signals, in order to extract 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 experiments in section 4 and finally we conclude this paper with remarks and future works in section 5.

2 Related works

While numerous tracking and recognition algorithms have been proposed in the computer vision community, these two topics were usually studied separately. For human face tracking, many different techniques have been developed, such as subspace-based methods [1], pixel-based tracking algorithms [2], contour-based tracking algorithms [3–5], and global statistics of color histograms [3, 6]. Likewise, there is a rich literature on face recognition published in the last 15 years [7, 8]; 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 [9], Zhou *et al.* [10] and Lee *et al.* [11] are the most relevant: all of them develop a tracking and recognition method using a unified probabilistic framework.

Our work is also closely related to the visual analysis of human motion, in particular with the automatic gait recognition (field of research). It is possible to classify the most important techniques in two distinct areas: holisitic approaches [12, 13], which aim to extract statistical features from a subject's silhouette in order to differentiate between individuals, and model-based approaches [14, 15], which aim to model human gait explicitly.

3 Recognition using head displacements

Our person recognition system is mainly composed by three parts: a video analyser for obtaining displacement signals, a feature extractor for computing feature vectors, and a person classifier for retrieving identities.

3.1 Video analyser module

The video analyser module 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 using a template matching technique in the RGB color space. The similarity measure is obtained by computing an Euclidean or city-block distance for each color component, then adding them (equal component weighting). If $\mathbf{T}_k$ is the actual template, $\mathbf{T}_{k-1}$ the previous one, $\mathbf{M}_{k-1}$ the latest match and α a weighting constant, then the template is updated with the following rule: $\mathbf{T}_k = \alpha \mathbf{M}_{k-1} + (1-\alpha)\,\mathbf{T}_{k-1}$. One can easily verify that the actual template is a weighted sum of all the previous ones and it can be set to include the limit cases of no update ($\alpha = 0$) and full update ($\alpha = 1$).

3.2 Feature extractor module

The feature extractor module deals with rough displacement signals of different head features, extracted from the video sequence.

In order to compute the feature vector, 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. By default, this module centers the signals and scales them, in order to remove any dependence on absolute head position and video resolution; it is also possible to impose an uniform variance, exploit polar coordinates or compute derivatives (like velocities or accelerations). It is important to notice that each signal has two components, usually the horizontal and vertical displacements, so the total number of different features F is the double of the number of face elements analyzed. In the following part, we are going to express all the feature vectors of a person q, extracted from his r-th video, with the following notation:

$$\mathbf{X}^{(q,r)} = \begin{bmatrix} \mathbf{x}_1^{(q,r)} \\ \vdots \\ \mathbf{x}_K^{(q,r)} \end{bmatrix}$$

where K is the total number of frames and $\mathbf{x}_k$ is a row vector representing the feature values computed from frame k.

3.3 Person recogniser module

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

The processed head displacements are used for training a Gaussian Mixture Model (GMM) for each person in the database, in order to model the characteristic displacements (or its derivatives) for that user; more precisely, the algorithm estimates the class-conditional probability density functions in a Bayesian classifier. Formally, the posterior probability for class ω_q is:

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

In our case, where each user has the same amount of videos, the priors and scaling factors are uniform and does not affect the posterior probability computation.

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

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

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, $\aleph\left(\mathbf{x}_k; \mu_c, \mathbf{\Sigma}_c\right)$.

It is important to underline that a part of the videos in the database is used for training those models, while the remaining sequences are used as tests for assessing the recognition performances (identification and verification scores).

4 Experiments and results

4.1 Data collection

Due to the lack of any standard video database for evaluating video person recognition algorithms, we collected a set of 144 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 6 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 1 for an example). Even though the videos are 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.

4.2 Experimental set-up

For our experiments, we selected 72 video sequences for training (8 for each of the 9 individuals), and the remaining 72 (out of 144) were left for testing. It is important to point out that there are no theoretical constraints concerning the number of videos per user and the total video chunks; on the other hand, it is necessary to have a few minutes for each individual for being able to learn the characteristical head motion and to train the individual GMMs.

We chose to extract the displacements of 4 head features - the eyes, nose and mouth - providing then 8 signals in total. During our experiments, we tested

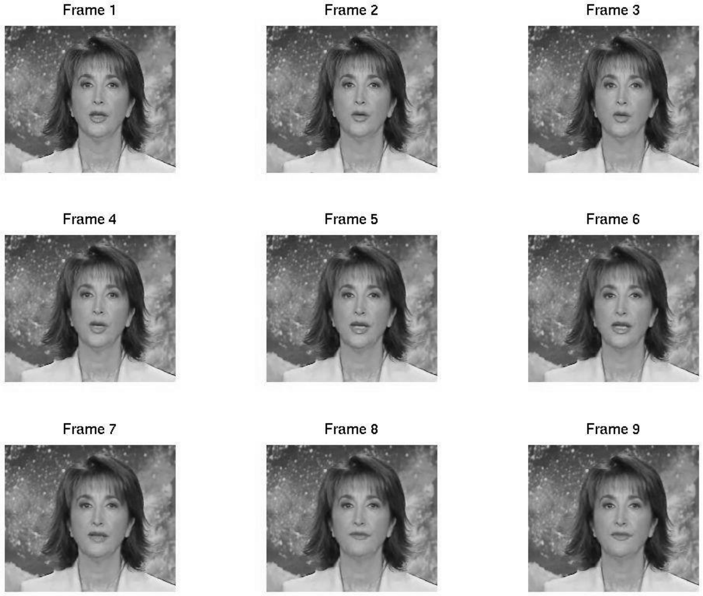

Fig. 1. The first 9 frames of a video sequence.

multiple configurations concerning the number of features to extract. The experimental results obtained using only two signals, like the eye displacements, were not as good as the actual choice of 4. We believe that even if our algorithm tracks the head in the image plane, the originating $3D$ movement - which can be represented with six parameters in the $3D$ space - needs more than just two points for proper estimation; moreover, more feature points may provide a higher precision for the global motion and can be sensible to small local deformations of the face (due to relative movement of the features).

As stated in section 3.1, the head detection part is done semi-automatically: the user must manually click on the (face) features of interest in the first frame, then then a tracking algorithm continues until the end of the sequence. After the localization of the head features - the eyes, nose and mouth - four templates of fixed size ($19 \times 25\, pixels$) are computed using the update formula. For the automatic tracking process, keeping the initial template ($\alpha = 0$) has showed the best discriminating properties, even if the algorithms is not always returning the correct match (absence of update); knowing the computational burden of a full template matching, we optimized the search window by taking into account the position of each feature and consequently analysing only small regions of the video frame ($74 \times 74\, pixels$).

Concerning the signal normalization, the most relevant results have been obtained using zero-mean; in fact, stronger constraints, like an uniform range or fixed variance, reduced the discriminating power and were abandoned. 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. We also tried to compute our feature vectors using first and second derivatives of the displacement signals - as velocities and accelerations - but the resulting recognition scores were not better than using only the original displacements.

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 [16], did not provide any advantage over the standard EM.

4.3 Identification and verification scores

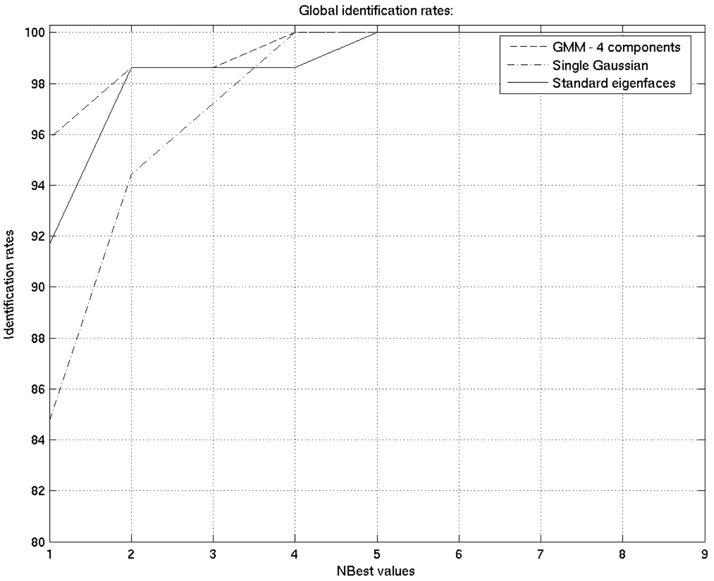

Fig. 2. Identification rates as a function of NBest values; for computing the scores, an individual is correctly identified if it is within the NBest matches.

Figure 2 shows the identification scores of our system: it is possible to notice that the identification rate is 95.8%, when considering the best match ($NBest = 1$), and 98.6%, when considering the three best matches ($NBest = 3$). Figure 3 shows the Receiver Operating Characteristic (ROC) curve of our system, with

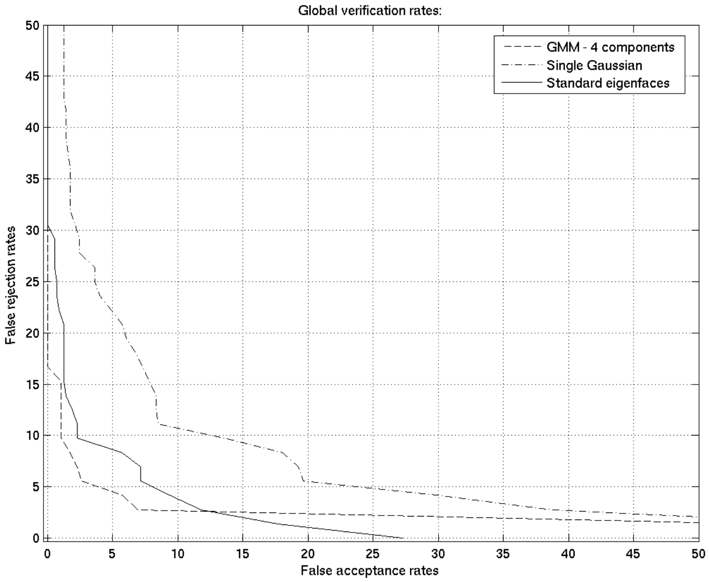

Fig. 3. Verification scores: False Rejection Rates (FRR) plotted as a function of False Acceptance Rates (FAR).

False Rejection Rates (FRR) plotted as a function of False Acceptance Rates (FAR): the Equal Error Rate (EER) value is 1.13%.

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 14 keyframes from each video chunk; on the other hand, only one keyframe was used in the testing phase. It's important to underline that the original keyframes have been manually normalized, by cropping the face region, then aligning and (in-plane) rotating 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 91.7%, rising up to 98.6% when considering the best three matches; the equal error rate of the system is 7.03%.

The previous experiments for recognising people from their head displacements are interesting; in fact, even if these signals could be considered as weak modalities and can not be as performing as the latest pixel-based techniques, they show that the behaviour of people can be a possible biometric. Moreover, our system is applied in real cases, with compressed video sequences and no constraints on movements or actions; our behavioural 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 sensible to within-subject variations: individuals may change their characteristic head motion when placed in different contexts or affected by particular emotional states.

4.4 Robustness to noise

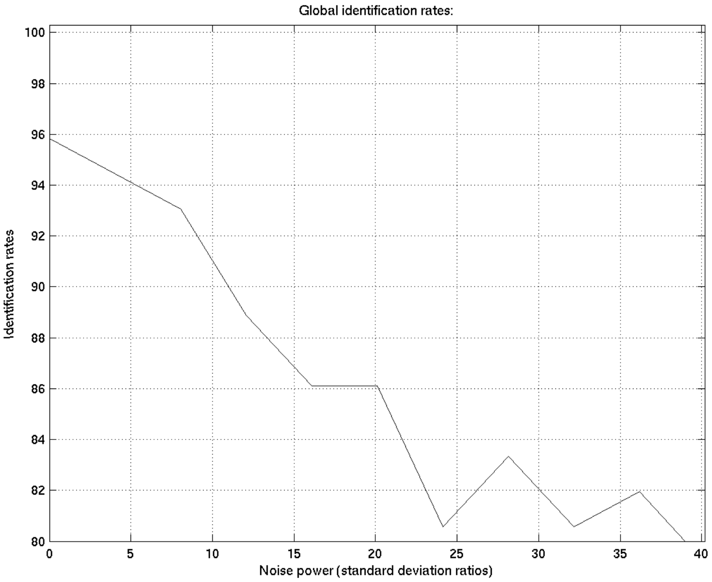

Fig. 4. Identification rates as a function of the noise power; in order to relate the energy of the signal with the one of the noise, the noise power is defined as the ratio between the standard deviation of the noise and the mean standard deviation of the signals (expressed in percentages).

In order to evaluate the robustness of our method from input noise, we artificially add a Gaussian noise with zero mean and variable standard deviation to all displacement signals (both training and testing sets), retrieved by the tracking module. In Figure 4, we report the identification scores as a function of the noise power; in order to relate the energy of the signal with the one of the noise, the noise power is defined as the ratio (expressed in percentages) between the standard deviation of the noise and the mean standard deviation of the signals. It is possible to notice that initially the identification rate is decreasing rapidly (−10% for noise power between 0 and 20%), then the loss is less important. One possible explaination may be the following: the characteristic small movements, which are really important for discriminating identities, are easily lost for low noise strenghts; after that the noise starts corrupting the global individual movements, less useful in recognition.

5 Conclusion and future works

This pioneering work on person recognition using head dynamics, retrieved in the image plane without the need of a complex $3D$ pose estimation, showed that the human behaviour and motion may be useful for discriminating people. Our study on head feature displacements represents a first step in the exploration of the face 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.

Our system can be improved by researching and implementing different solutions. One way is to use our biometric system, based on head displacements, and integrate it in a multimodal one; for this purpouse it could be possible to couple it with a physical modality, like the appearance of the face, or with another behavioural modality, like eye blinking or lip movements. Considering the low quality of our video database, in which fine details are affected by compression noise, the former case seems more feasible. Another possibility is to refine the signal extraction process, implementing a more robust tracking algorithm than the RGB template matching. Although it is reasonable that more precise signals could provide better classification power, the quality of those already extracted is actually good enough for our algorithm; in fact, by manual inspection we noticed that the tracking points are almost always correct (considering a pixel of tolerance) and that the occasional errors have reduced influence on our statistical approach. It may be also interesting to focus the analysis on individual gestures and exploit that knowledge for classifying identities; as an example, the head dynamics might be analysed in a local way, computing feature vectors in each temporal window. This approach may show more important discriminating power, capturing the details of personal movement, but the lack of prior information on the evolution of the motion and the relatively small size of the training database could be overwhelming. Finally, all our identification and verification results should be validated on a bigger database.

Acknowledgment

The authors would like to thank the anonymous reviewers for their useful comments.

References

1. G. D. Hager and P. N. Belhumeur, "'Efficient region tracking with parametric models of geometry and illumination"', in *Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, October 1998.

2. A. D. Jepson, D. J. Fleet and T. R. El-Maraghi, ”‘Robust online appearance models for visual tracking”’, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, 8–14 December 2001, Toronto, Canada, vol. 1, pp. 415–422.
3. S. Birchfield, ”‘Elliptical head tracking using intensity gradients and color histograms”’, in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 23–25 June 1998, Stanford, USA, pp. 232–237.
4. Y. Chen, Y. Rui and T. S. Huang, ”‘JPDAF based HMM for real-time contour tracking”’, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, 8–14 December 2001, Toronto, Canada, vol. 1, pp. 543–550.
5. Y. Wu and T. S. Huang, ”‘A co-inference approach to robust visual tracking”’, in *Proceedings of the Eighth IEEE International Conference on Computer Vision (ICCV 2001)*, 7–14 July 2001, Urbana, USA, vol. 2, pp. 26–33.
6. D. Comaniciu, V. Ramesh and P. Meer, ”‘Kernel-based object tracking”’, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. P. S. Huang, C. J. Harris and M. S. Nixon, ”‘Recognising humans by gait via parametric canonical space”’, in *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 359–366, October 1999.
13. J. B. Hayfron-Acquah, M. S. Nixon and J. N. Carter, ”‘Automatic gait recognition by symmetry analysis”’, in *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2175–2183, September 2003.
14. D. Cunado, M. S. Nixon and J. N. Carter, ”‘Automatic extraction and description of human gait models for recognition purposes”’, in *Computer Vision and Image Understanding*, vol. 90, no. 1, pp. 1–41, April 2003.
15. C. Yam , M. S. Nixon and J. N. Carter, ”‘Automated person recognition by walking and running via model-based approaches”’, in *Pattern Recognition*, vol. 37, no. 5, pp. 1057–1072, May 2004.
16. 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.