Towards Person Recognition Using Head Dynamics

Federico MATTA
Eurécom Institute
2229 route des Cretes, BP 193
06904 Sophia Antipolis, FRANCE
Federico.Matta@eurecom.fr

Jean-Luc DUGELAY
Eurécom Institute
2229 route des Cretes, BP 193
06904 Sophia Antipolis, FRANCE
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 pixel-based information, our approach exploits the head dynamics; in particular the displacement signals of few head features. 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 preliminary approach are nevertheless encouraging.

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 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 pixel-based information, such as greylevel values. While these systems have indeed produced very low error rates, they ignore other levels of information that can be used for discriminating identities; the analysis of human motion and, more precisely, facial mimics or head dynamics, may represent a valid biometry and may be used as an alternative, or jointly, with pixel-based techniques and other biometrics.

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 global motion information, and used for discriminating identities with a nearest neighbour 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 vision community, these two topics were usually studied separately. For human face tracking, many different techniques have been developed, such as subspace-based methods [8], pixel-based tracking algorithms [1], contour-based tracking algorithms [3, 15, 14], and global statistics of color histograms [3, 7]. Likewise, there is a rich literature on face recognition published in the last 15 years [4, 13]; 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 to each frame, without taking into the 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 [2], Zhou, Krueger & Chellappa [12] and Lee, Ho, Yang & Kriegman [10] are the most relevant: all of them develop a tracking and recognition method using a 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 [11, 9], which aim to extract statistical features from a subject's silhouette to differentiate between subjects, and model-based approaches [6, 5], 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; 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 person classifier module deals with rough displacement signals of different head features, extracted from the video sequence. Firstly, the system centers the signals and optionally scales them, in order to have a uniform range or variance; after that, a feature vector is computed from the displacement values and used in the classification task. It is important to notice that for each of the F features there are two signals, usually the horizontal and vertical displacements. More formally, if $\mathbf{s}_i = [s_{i,1}, \ldots, s_{i,K}]$ is a zero-mean unidimensional displacement signal i, then it is possible to represent them all with the following matrix notation: $\mathbf{S} = [\mathbf{s}_1^t, \ldots \mathbf{s}_{2F}^t]$. The algorithm then computes the covariance matrix of the signals, $\mathbf{C} = \mathrm{cov}\,(\mathbf{S})$, and forms the feature vector using its values; this process corresponds to model the head displacements using a $2F$ dimensional Gaussian probability function, and to extract its shape parameters.

3.3. Person recognition module

The last module exploits the feature vectors computed from video sequences for classification purposes. One part of the video database is used for calculating the identity models in the feature space, while the remaining sequences are used as tests for assessing the recognition performances (identification and verification scores). A nearest neighbour classifier is implemented for discriminating the identities: it computes the Euclidean distance between identity models and test (feature) vectors and it returns the nearest pairs.

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 91 video sequences of 7 different persons, for the task of 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 320×240 $pixels$ and a temporal resolution of 30 $frames/second$, and lasts 10 $seconds$ (refer to Figure 1 for an example). Even though the videos are low quality, compressed at 300 $Kbits/second$, 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.

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

4.2. Experimental set-up

For our experiments, we selected 35 video sequences for training (5 for each of the 7 individuals), and the remaining 56 (out of 91) were left for testing. We chose to extract the displacements of 4 head features, the eyes, nose and mouth, providing then 8 signals in total (refer to Figure 2 for an example). For the tracking process, keeping the initial template ($\alpha = 0$) has showed the best discriminating properties, even if the process in 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 only; in fact, stronger constraints, like a uniform range or variance, reduced the discriminating power and were abandoned. It is important to notice that all the videos have almost equal head sizes and zooms, so there is no need for spatial scal-

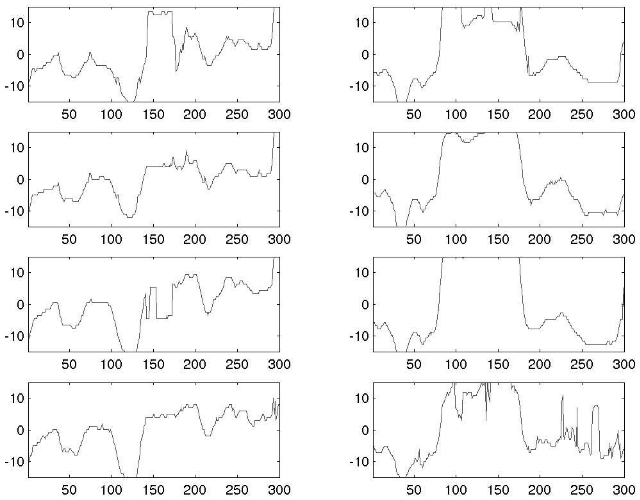

Figure 2. Displacement signals for (top-down) left eye, right eye, nose and mouth; the first column represents the horizontal displacements, while the second the vertical ones.

ing.

In the classification part, each identity model has been computed by taking the mean of 5 training feature vectors, belonging to the same individual; this simple strategy showed important robustness to intra-personal variability in the training set. During our experiments we also tested our classifier with frequency-based feature vectors (using spectral energy, for example), but the discriminative power was definitively less important.

4.3. Identification and verification scores

Figure 3 shows the identification scores of our system: it is possible to notice that the identification rate is 62%, when considering the best match ($NBest = 1$), and 80%, 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 nearly 0.32.

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 following results have been obtained with normalized grey images, centered and resized to $64 \times 64 pixels$, and an eigenspace of dimension 30. The identification rate for the best match is 93%, rinsing up to 100% when considering the best three matches; the equal error rate of the system is 0.13. Finally, an algorithm which randomly matches inputs with identities, would obviously obtain an identification rate of 14%, when retaining the best candidate.

The previous experiments for recognising people from their head displacements are encouraging; in fact, even if these signals could be considered as weak modalities

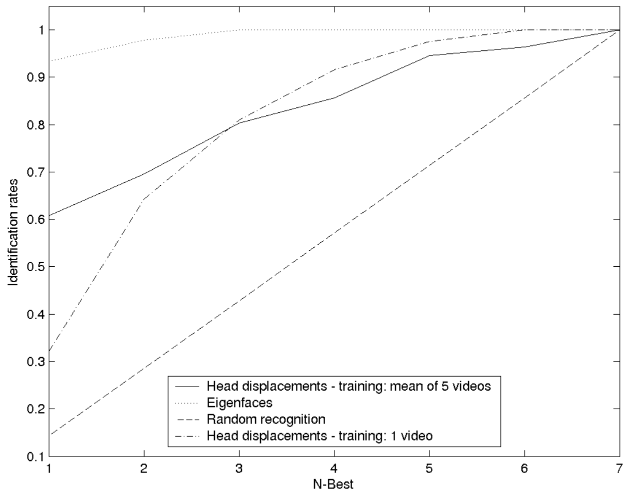

Figure 3. Identification rates as a function of NBest values; for computing the scores, an individual is correctly identified if it is within the NBest matches.

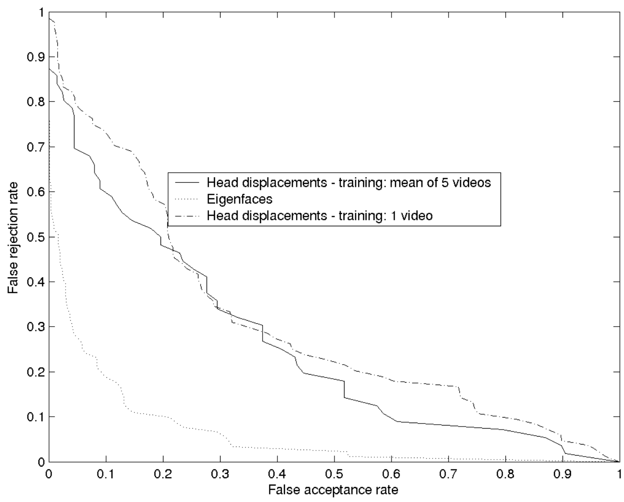

Figure 4. Verification scores: False Rejection Rates (FRR) plotted as a function of False Acceptance Rates (FAR).

and can not be as performing as static pixel-based techniques, they show that the behaviour of people may be a valid 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.

5. Conclusion and future works

This pioneering work on person recognition using head dynamics showed that the human behaviour and motion may be useful for discriminating people, even if this preliminary results are not as good as state of the art pixel-based techniques. Our study on head feature displacements represents the first step in the exploration of the face dynamics and their potential use in recognition: instead of considering the physical aspects of the face, our purpose is to consider the behavioural factors and their possible applications.

Our system can be improved by researching and implementing different solutions. One way is to refine the signal extraction process, probably developing a more robust tracking algorithm than the RGB template matching; more precise signals could logically provide better recognition results. Another possibility is to research more discriminating features, either pursuing our global (statistical) approach, either trying to analyse the head dynamics punctually (using temporal windows, for example); this second choice may show more important discriminating power, capturing the details of personal movement, but the absence of constraints and the lack of prior information on the evolution of the motion can be overwhelming. Finally, assuming the availability of a larger training database, the recognition results can be improved with a more sophisticated classifier, than our nearest neighbour.

On the other hand, this system can be coupled with a state of the art pixel based technique, or developed for being able to analyse different behaviours, like eye blinking our mouth dynamics, and use them for recognising people. In fact, even if each single modality may not provide a superior discriminating power, a system exploiting multiple face dynamics is willing to obtain important results.

References

[1] T. F. E.-M. Allan D. Jepson, David J. Fleet. Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1296–1311, October 2003.

[2] R. C. Baoxin Li. A generic approach to simultaneous tracking and verification in video. *IEEE Transactions on Image Processing*, 11(5):530–544, May 2002.

[3] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. *Proceedings IEEE on Computer Vision and Pattern Recognition*, pages 232–237, June 1998.

[4] S. S. Charles L. Wilson. Human and machine recognition of faces: a survey. *Proc. IEEE*, 83(5):705–741, May 1995.

[5] J. N. C. ChewYean Yam, Mark S. Nixon. Automated person recognition by walking and running via model-based approaches. *Pattern Recognition*, 37(5):1057–1072, May 2004.

[6] J. N. C. David Cunado, Mark S. Nixon. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90(1):1–41, April 2003.

[7] P. M. Dorin Comaniciu, Visvanathan Ramesh. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003.

[8] P. N. B. Gregory D. Hager. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, October 1998.

[9] J. N. C. James B. Hayfron-Acquah, Mark S. Nixon. Automatic gait recognition by symmetry analysis. *Pattern Recognition Letters*, 24(13):2175–2183, September 2003.

[10] M.-H. Y. D. K. Kuang-Chih Lee, Jeffrey Ho. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, April 2005.

[11] M. S. N. Ping S. Huang, C. J. Harris. Recognising humans by gait via parametric canonical space. *Artificial Intelligence in Engineering*, 13(4):359–366, October 1999.

[12] R. C. Shaohua Zhou, Volker Krueger. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(1-2):214–245, July-August 2003.

[13] P. J. P. A. R. Wen-Yi Zhao, Rama Chellappa. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[14] T. S. H. Ying Wu. A co-inference approach to robust visual tracking. *Proceedings IEEE on Computer Vision*, 2:26–32, July 2001.

[15] T. S. H. Yunqiang Chen, Yong Rui. Jpdaf based hmm for real-time contour tracking. *Proceedings IEEE on Computer Vision and Pattern Recognition*, 1:543–550, December 2001.