TOMOFACES: EIGENFACES EXTENDED TO VIDEOS OF SPEAKERS

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 propose a novel spatio-temporal approach for person recognition using video information. By applying discrete video tomography, our algorithm summarises the head and facial dynamics of a sequence into a single image (called "video X-ray image"), which is subsequently analysed by an extended version of the eigenface approach. In the experimental part, we assess the discriminative power of our system and we compare it with an analogous one working on traditional facial appearance. Finally, we integrate the X-ray information with appearance in a multimodal system, which improves the recognition rates of standalone frameworks.

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 [1], 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.

Person recognition using videos has some advantages over image-based recognition. Firstly, the temporal information can be exploited to facilitate the recognition task: for example, the global head motion and the local facial motion provide additional cues to the characterisation of the individual, and can be used as biometric identifiers. Secondly, 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 enables learning or updating subject models over time.

Nevertheless, most of video-based approaches proposed in the literature are straightforward extensions of image-based algorithms, and exploit the video as a source of data, neglecting the temporal information. Instead, we aim to use the whole spatio-temporal information for person recognition; for this reason, we propose a novel approach that summarises the head and facial dynamics of a video clip into a single image (called "video X-ray image"), which is subsequently analysed by an extended version of the eigenface [2] approach.

The article is organised as follows. In Section 2 we briefly review the related works proposed in the research literature; then, in Section 3 we describe the recognition system using X-ray images. After that, in Section 4 we illustrate the experimental set-up and results and finally in section 5 we conclude our article with final remarks and perspectives.

2. RELATED WORKS

Eigenfaces [2] is one of the essential basic techniques for person recognition by using facial appearance. It has been widely studied and largely applied to image data; for a detailed review, the interested reader can refer to [1]. In the scientific literature, there are a few extensions of the standard image-based approach to video data, mostly exploiting the additional spatial information of the video and not fully exploiting the temporal one. Some of these extensions adopt a straightforward multiframe strategy, in which decision fusion techniques integrate the opinions on each frame. In particular, Satoh [3] considered the smallest distance between frame pairs and Huang and Trivedi [4] applied the majority voting rule or a post-classifier (based on discrete HMMs) on individual opinions. There are other strategies which exploit the abundant video data to train statistical models of the individual facial manifolds. They represent applications of the subspace method and its variants: we cite the self-eigenface approach of Torres and Vila [5], the CLAFIC-based methods of Satoh [3] and the mutual subspace method of Yamaguchi et al. [6] and Nishiyama et al. [7].

Discrete video tomography was introduced by Akatsu and Tonomura [8] for camera work analysis, which is the estimation of camera motion parameters (panning, tilting, zooming...) in video sequences. In the original approach,

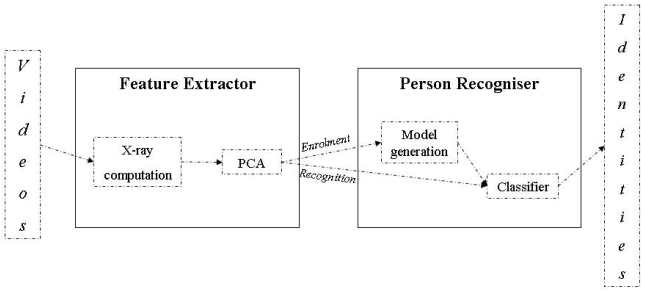

Fig. 1. Block diagram of the recognition system.

Fig. 2. Example of temporal X-ray transformation. From left to right, starting from the top: original frame, edge image, temporal X-ray image, filtered temporal X-ray image.

the "video X-ray images", that are spatio-temporal images obtained by computing the average of each row or column in successive frames, are generated after an edge detection step. In a successive work, Joly and Kim [9] proposed a computationally reduced algorithm, using the intensity information and avoiding the edge detection calculation.

3. DESCRIPTION OF THE SYSTEM

Our person recognition system is an extension of the eigenface approach [2] to "video X-ray images". As shown in Figure 1, it is composed by two modules: a Feature Extractor, which transforms input videos into X-ray images and extracts low dimensional feature vectors, and a Person Recogniser, which generates user models for the client database (enrolment phase) and matches unknown feature vectors with stored models (recognition phase).

3.1. Feature Extractor

Inspired by the application of discrete video tomography [8] for camera motion estimation, we compute the temporal X-ray tranformation of a video sequence, to summarise the head motion information of a person into a single X-ray image. It is important to notice that we restrict our framework to a fixed camera and background; hence, the video X-ray images represent only the motion of the head, which is the information that we use to discriminate identities.

Given an input video of length T_i, $V_i \equiv \{I_{i,1}, \ldots, I_{i,T_i}\}$, the Feature Extractor module firstly calculates the edge image sequence, E_i, obtained by applying the Canny edge-finding method [10] frame by frame:

$$E_i \equiv \{J_{i,1}, \ldots, J_{i,T_i}\} = f_{EF}(V_i) \tag{1}$$

Then, the resulting binary frames, $J_{i,t}$, are temporally added up to generate the X-ray image of the sequence:

$$X_i = C \sum_{t=1}^{T_i} J_{i,t} \tag{2}$$

where C is a scaling factor to adjust the upper range value of the X-ray image.

Figure 2 presents a visual example of the steps described above. By looking at the lower left picture, corresponding to the X-ray image, it is possible to notice that the static textured background generates very dark areas and very vivid contours; this information is not related to the personal motion and may negatively affect the discriminative power of the image. For this reason, the Feature Extractor filters the X-ray image in order to attenuate its brightest background contours, by putting to black all the pixels above a threshold value.

After that, the Feature Extractor reduces the X-ray image space to a low dimensional feature space, by applying the principal component analysis (PCA) (also called the Karhunen-Loeve transform (KLT)): PCA computes a set of orthonormal vectors, which optimally represent the distribution of the training data in the root mean squares sense. In the end, the optimal projection matrix, $\mathbf{P}$, is obtained by retaining the eigenvectors corresponding to the M largest eigenvalues, and the X-ray image is approximated by its feature vector, $\mathbf{y}_i \in \Re^M$, calculated using the following linear projection:

$$\mathbf{y}_i = \mathbf{P}^T (\mathbf{x}_i - \mu) \tag{3}$$

in which $\mathbf{x}_i$ is the filtered X-ray image in a vectorial form and μ is the mean value.

3.2. Person Recogniser

During the enrolment phase, the Person Recogniser module generates the client models and stores them into the system. These representative models of the users are the cluster centres in feature space that are obtained using the enrolment data set.

For the recognition phase, the system implements a nearest neighbour classifier which compares unknown feature vectors with client models in feature space. The similarity

Fig. 3. Examples of variations in our video database: different backgrounds, aging effects, various clothings and haircuts.

measure adopted, $S(\cdot)$, is inversely proportional to the cosine distance:

$$S(\mathbf{y}_i, \mathbf{y}_j) = 1 - \frac{{\mathbf{y}_i}^T \mathbf{y}_j}{\|\mathbf{y}_i\| \, \|\mathbf{y}_j\|} \tag{4}$$

and has the property to be bounded into the interval $[0, 1]$.

4. EXPERIMENTS AND RESULTS

4.1. Video database

Unfortunately, the existing standard video databases (like ValidDB or XM2VTSDB) do not match the requirements for efficiently testing the proposed algorithm; in fact, they are generally intended for physiological or multimodal approaches, and they do not present observable and characteristic motion, nor enough video recordings per user to allow the development of our behavioural approach.

For this reason, we have been collecting a set of 208 video sequences belonging to 13 different persons, and we have trained and tested our system with this data. The video chunks present TV speakers announcing the news of the day: their behaviour and motion are natural, without any constraint imposed to their movement, pose or action. A typical sequence has a spatial resolution of 352×288 $pixels$ and a temporal resolution of 23.97 $frames/second$, and lasts almost 14 $seconds$. The videos are of low quality, acquired using a fixed camera and compressed at 300 $Kbits/second$ (including audio), and they have been collected during a period of 18 months. All these elements contribute to create a rich and realistic database, with multiple variations related to: different backgrounds, aging effects, various clothings and haircuts and presence/absence of glasses and beard; some visual examples of variations are depicted in Figure 3.

4.2. Experimental set-up

For our experiments, we split the database into two distinct sub-sets: one with 8 $videos/person$ for the enrolment, and another one with the remaining 8 $videos/person$ for the recognition tests. In total, each sub-set contains 104 $sequences$, 34320 $frames$ and almost 24 $minutes$ of video data.

Method	CIR (1st) (%)	CIR (3rd) (%)	EER (%)
X-ray	75.00	96.15	7.69
Appearance	72.12	96.15	8.25

Table 1. Identification and verification results for individual approaches (X-ray and facial appearance).

Method	CIR (1st) (%)	CIR (3rd) (%)	EER (%)
Equal w.	80.77	100.00	4.89
Adaptive w.	78.85	100.00	5.89

Table 2. Identification and verification results for integrated recognition systems, using equal weighting (mean) and adaptive weighting.

The recognition system has been tested using a feature space of size 81 constructed with the enrolment data set. The video frames are also pre-processed doing an histogram equalisation, in order to reduce the illumination variations between different sequences. Concerning the attenuation of the background, we filtered all the pixels above 66% of the grey level range in the X-ray images.

4.3. Comparison

We compare the results obtained using X-ray images with those using traditional pictures of facial appearance. By replacing the X-ray computational step with a video frame processor in Figure 1, we convert our recognition system into one similar[1] to the original eigenface approach [2]. This video frame processor generates an image database derived from the video database depicted before: after histogram equalisation, it subsamples each video in the enrolment set at 2 $frames/second$, thus extracting 28 $frames/video$, while for the testing set it only retrieves the first keyframe.

4.4. Recognition results

The identification and verification results for the two recognition systems are summarised in Table 1; its columns report the correct identification rates (CIR), computed using the best and 3-best matches, and the equal error rates (EER) for the verification operational mode. We notice that the recognition system based on X-ray images performs better than the analogous one working with facial appearance.

In Table 2 we also present the results of two recognition systems, which integrate the information from the X-ray and facial images. More precisely, we operate a fusion of the individual similarity scores, by using the weighted sum rule:

[1] In fact, our system is working with colour images and using the cosine distance, instead of greyscale pictures and Euclidean distance.

in one case we consider an equal weighting (by calculating the mean of the scores), in the other one we implement the adaptive weighting rule proposed by Chang et al. [11]. Both integration schemes perform closely, and clearly improve the recognition results of the individual modalities, shown in Table 1.

It is important to notice that the algorithms have been tested using video frames with no spatial normalisation. In fact, our framework considers a real case with actual videos, and try to avoid the need of a high quality normalisation step, which is hard to achieve in practice in an automatic way. On the other hand, due to the well known high sensitivity of PCA-based recognition algorithms to facial alignment and variations in pose and scale, most of the systems proposed in the literature are tested with normalised images. In the experiments done with an "ad-hoc" configuration, using perfectly manually normalised frames[2] and no additional noisy information from background and clothing, the eigenface-like system achieves perfect recognition in both identification and verification tasks.

5. CONCLUSION AND FUTURE WORKS

The major contribution of our study is that, as far as we know, this is the first attempt to employ "video X-ray images" for person recognition. In fact, in this article we propose a novel spatio-temporal approach and we experimentally verify its discriminative power. As a secondary contribution, we show that it can be directly integrated with a recognition algorithm using facial appearance, and that the combined system achieves better performances. This method has also the advantage of not requiring complex pre-processing, like spatial normalisations of frames or temporal synchronisations of video chunks.

Concerning future works, first of all we are aware that our results must be verified by a bigger experimental validation. Then, our system could be improved by using another subspace reduction technique instead of PCA, like linear discriminant analysis (LDA) for example. Finally, before its practical application to video surveillance or access control tasks, we need to evaluate the effect of uncontrolled body and camera motion in the X-ray computation, and probably refine our strategy to compensate for it.

6. REFERENCES

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

[2] Turk M.A. and Pentland A.P., "Face recognition using eigenfaces.," *IEEE Proceedings on Computer Vision and Pattern Recognition*, pp. 586–591, June 1991.

[3] Satoh S., "Comparative evaluation of face sequence matching for content-based video access.," *IEEE Proceedings on Automatic Face and Gesture Recognition*, pp. 163–168, March 2000.

[4] Huang S.K. and Trivedi M.M., "Streaming face recognition using multicamera video arrays.," *Proceedings of Pattern Recognition*, vol. 4, pp. 213–216, 2002.

[5] Torres L. and Vila J., "Automatic face recognition for video indexing applications.," *Pattern Recognition*, vol. 35, no. 3, pp. 615–625, March 2002.

[6] Yamaguchi O., Fukui K., and Maeda K.I., "Face recognition using temporal image sequence.," *IEEE Proceedings on Automatic Face and Gesture Recognition*, pp. 318–323, April 1998.

[7] Nishiyama M., Yamaguchi O., and Fukui K., "Face recognition with multiple constrained mutual subspace method.," *Proceedings of Audio- and Video-Based Biometric Person Authentication*, vol. 3546/2005, pp. 71–80, June 2005.

[8] Akutsu A. and Tonomura Y., "Video tomography: an efficient method for camerawork extraction and motion analysis.," *Proceedings on ACM Multimedia*, pp. 349–356, October 1994.

[9] Joly P. and Hae-Kwang K., "Efficient automatic analysis of camera work and microsegmentation of video using spatio-temporal images.," *Signal Processing: Image Communication*, vol. 8, no. 4, pp. 295–307, May 1996.

[10] Canny J., "A computational approach to edge detection.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[11] Chang K.I., Bowyer K.W., and Flynn P.J., "An evaluation of multimodal 2d+3d face biometrics.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 619–624, April 2005.

[2] The perfect manual normalisation consists of: cropping the face region, then aligning and in-plane horizontally rotating the heads.