4 Septembre 2006

SP2 : Segmentation et Authentification du visage

Etat de l’art sur l’utilisation de la dimension temporelle en reconnaissance de visage

F.Matta, U.Saeed, C.Mallauran, J-L.Dugelay

Avant-Propos

La reconnaissance de visage est un domaine de recherche très actif. La plupart des algorithmes font appel aux images fixes pour identifier une personne. Dans sa thèse, F.Perronnin [13] propose un état de l'art sur la reconnaissance de visage en images fixes. Cependant, les approches basées image présentent quelques faiblesses, dues notamment aux variations d'expression, de pose et de conditions en luminosité. C'est pourquoi, les chercheurs se sont tournés vers la vidéo, qui présente quelques avantages non négligeables. A l'heure actuelle, l'approche la plus utilisée considère la vidéo comme une séquence d'images fixes. Une première consiste à appliquer l'algorithme sur plusieurs images issues de la vidéo et réaliser ensuite une fusion des scores. Une autre méthode consiste à sélectionner une image parmi celles disponibles en vue d'obtenir les meilleurs résultats possibles lors de la reconnaissance. Néanmoins, toutes ces méthodes ne tirent pas avantage des informations temporelles contenues dans la vidéo. D'ailleurs, actuellement, peu de recherches portent sur ce sujet. Ce document a donc pour objectif de présenter les quelques approches exploitant l'aspect dynamique de la vidéo.

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, where both the training and test set 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. First, the temporal information of faces can be exploited to facilitate the recognition task; for example, person specific dynamic characteristics like the motion of the head, the evolution of the pose or the mimic of the face. Second, 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 allows learning or updating the subject model over time.

Currently, most face recognition techniques using videos are straightforward generalizations of image face recognition algorithms: in these systems, the individual image recognition strategy is applied independently for each frame, without taking into account the temporal information enclosed in the sequence. However, a few recent attempts propose new approaches for person recognition, aiming to exploit the temporal information or the human behaviour embedded in video sequences.

In the next subsection we will summarize those approaches, recently proposed to the scientific community.

1 Strategies exploiting temporal information in videos

Person recognition systems that exploit the video temporal information are recently emerging to the recognition 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 of classification. We then propose the following categorization:

- *Holistic methods*. This family of techniques analyse the head as a whole, by extracting the head displacements or the pose evolution. In our particular case, we can finely discriminate temporal approaches by dividing those which are based on pose evolution and trajectory from those which track and recognise in a unified framework.
- *Feature-based methods*. These methods typically exploit the individual facial features, like the eyes, nose, mouth and eyebrows.
- *Hybrid methods*. Hybrid approaches use both holistic and local features.

1.1 Holistic methods

In [3], Li et al. propose an approach to model-based dynamic object verification and identification using videos. They first compute the motion trajectory of the object, then they estimate its 3D pose at each time step, searching on a space constrained by the trajectory information and scene structure; finally, the matching part for estimation and recognition is solved using a generalized Hausdorff metric on edge maps. As indicators of confidence for the identification/verification tasks, the authors also introduce two measures based on the smoothness of the pose evolution curve and on the concentration of the band of pose estimates. A few results are provided for infrared and optical sequences, but no systematic evaluation of recognition is done.

In 2002, Li and Chellappa [4] were the first to develop a generic approach to simultaneous object tracking and verification in video data, using posterior probability density estimation through sequential Monte Carlo methods. Visual tracking is formulated as a probability density propagation problem, with the density being defined over a proper state space characterizing the object configuration. In the end, verification is realized through hypothesis testing using the estimated posterior density. In this work, Li and Chellappa present a noticeable probabilistic framework that has been successfully applied in vehicle, human and face tracking and verification; again, no systematic evaluation of recognition is done.

Huang and Trivedi in [10] describe a multi-camera system for intelligent rooms, combining PCA based subspace feature analysis with Hidden Markov Models (HMM). Video streams containing several faces are first partitioned into overlapping and non overlapping sequences of fixed length, and then a modified PCA type analysis is applied to each frame separately in which vectors are normalized. Three classification schemes are applied to the videos. In the first one each frame is classified and a majority rule is applied to the entire sequence. In the second scheme a Discrete HMM is created by using several training sequences for each person using Baum-Welch estimation, then a test sequences is classified by the forward method. In the last scheme the feature vectors from the PCA analysis are used directly with a Continuous Density HMM and classification is achieved by a maximum likelihood using forward method. Experiments were carried out on the NOVA system using 5 training and 4 testing videos for each of the 5 subjects. Following results were reported; 81.7 % identification rate for the first scheme, 88.7 % for the second and 99.0 % for the third scheme.

Following the same principle, Zhou et al. [5] propose a similar probabilistic framework, where a time series state space model is applied to fuse temporal information. More precisely, the joint posterior distribution of the motion vector and the identity variable is estimated at each time step and then propagated to the next time step. Compared to the work of Li and Chellappa [4], their model includes the identity variable in the parameterization of the state space model, providing a more general framework and a more robust estimation of the posterior distribution of the identity variable, through marginalization over the motion vector. In their work, Zhou et al. also develop an exemplar-based learning strategy to automatically select video representatives, serving as mixture centres in an updated likelihood measure.

Liu and Chen [6] propose a recognition system based on adaptive Hidden Markov Models (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. Then, in the classification part, identification is achieved by using the likelihood scores provided by the HMMs; finally each HMM is automatically adapted with the test video sequence, which results in better modelling over time.

In [11] Aggarwal et al. have modeled the moving face as a linear dynamical system using an autoregressive and moving average (ARMA) model. The system starts by a preprocessing step which crops the face to a fixed size, then the nose tip is tracked by a KL tracker and pose is estimated by a rough edge based technique. Next the parameters of the ARMA model are estimated for the entire database using the closed form solution. Finally the recognition is performed by calculating the subspace angles between the ARMA models of the gallery and the probe images. The experiments were performed on 2 databases, for both of the databases the subjects were arbitrarily moving their heads and expressions, the lighting conditions were also quite different. A recognition rate of 90 % was achieved on both the databases.

Recently, Lee et al. [7] 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, a low dimensional piecewise linear approximation of the appearance manifold. The connectivity among the subspaces is represented by a transition matrix, which is directly learned from training sequences by observing the actual transitions between pose states. Finally, recognition is achieved through Bayesian inference and a maximum likelihood estimate.

In [8], Matta and Dugelay propose a new recognition system based on head motion, which exploits the temporal information and the human behaviour embedded in video sequences. By analysing head displacements, the authors try to capture the characteristic movement of each individual, and then to exploit this information for discriminating identities. Head motion is firstly analysed by retrieving the displacements of the eyes, nose and mouth in each video frame; then, the raw signals are transformed and normalized in order to obtain video independent feature vectors. Finally, in order to extract the behavioural information related to each individual and use it for identification and verification purposes, they train individual Gaussian Mixture Models (GMMs) and achieve classification through a Bayesian classifier.

1.2 Feature-based methods

One of the first attempts to exploit facial motion for identifying people is presented by Chen et al. in [2]. In their work, they propose to use the optical flow extracted from the motion of the face for creating a feature vector used during identification. More precisely, they concatenate the optical flows belonging to each frame into a high dimensional feature vector, which is subsequently reduced using a Principal Component Analysis (PCA) followed by a Linear Discriminant Analysis (LDA). This approach demonstrates robustness to illumination variations, but in general its performances are not as good as state-of-the-art face recognition techniques. The main drawback of the algorithm is in the direct concatenation of the optic flows, which leads to a synchronization issue if the model and test video have different motions.

1.3 Hybrid methods

Colmenarez et al. in [12] 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. The face is divided into 4 feature region images containing 9 facial features which are detected and tracked automatically. The feature region images are modeled using Gaussian distributions on principal component subspace. Experiments were carried out on a small database of 18 people with 3 video sequences each. The training was supervised and used the first 2/3rd frames of the first sequence which were labeled by hand. Only relative results were provided for face recognition whereas for expression recognition the error varied form 6.8 % to 20.5 % for various expressions.

Recently in [9], Saeed et al. have augmented the feature vector of a previous system [8] with features extracted from the motion of the mouth. The exploit the rough localization of the mouth provided by the head feature extractor and develop a simple algorithm based on a color transform that enhances the mouth and then detects the edges using Sobel edge detector with Otsu's thresholding, followed by a series of morphological operations to improve the shape of the detected mouth. Then features such as the area contained in lip, major / minor axis of lip are extracted and arranged in the feature vector and used for recognition.

References

[1] Turk M., and Pentland A. Eigenfaces for Recognition, in *Journal of Cognitive Neuroscience*. Vol. 3, no. 1, pp. 71-86, 1991.

[2] Chen L., Liao H., and Lin J. Person identification using facial motion, in *Proceedings of International Conference on Image Processing*, Thessaloniki, Greece, 7-10 October 2001, pp. 677-680.

[3] Li B., Chellappa R., Zheng Q., and Der S. Model-based temporal object verification using video, in *IEEE Transactions on in Image Processing*. Vol. 10, no. 6, pp. 897-908, June 2001.

[4] Li B., and Chellappa R. 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.

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

[6] Liu X., and Cheng T. Video-based face recognition using adaptive hidden Markov models, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 18-20 June 2003, vol. 1, pp. 340-345.

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

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

[9] Saeed U., Matta F. and Dugelay J-L. Person Recognition based on Head and Mouth Dynamics, in *Proceedings of International Workshop of Multimedia Signal Processing (MMSP2006)*. 3-6 October 2006, Victoria, BC, Canada.

[10] Huang K. S., and Trivedi M. M. Streaming Face Recognition using Multicamera Video Arrays, in *Proceedings of 16th International Conference on Pattern Recognition*. 11-15 August 2002, vol. 4, pp. 213-216.

[11] Aggarwal G., Chowdhury A. K. R., and Chellappa R. A System Identification approach for Video-based Face Recognition, in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR2004*). 23-26 August 2004, vol. 4, pp. 175-178.

[12] Colmenarez A., Frey B., and Huang T. S. A Probabilistic Framework for Embedded Face and Facial Expression Recognition, in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 23-25 June 2003, vol. 1, pp. 592-597.

[13] Perronnin F., A probabilistic model of face mapping applied to person recognition, Ph. D. Thesis, 2004