Person Authentication from Video of Faces: A Behavioral and Physiological Approach Using Pseudo Hierarchical Hidden Markov Models

Manuele Bicego[1], Enrico Grosso[1], and Massimo Tistarelli[2]

[1] DEIR - University of Sassari, via Torre Tonda 34 - 07100 Sassari - Italy
[2] DAP - University of Sassari, piazza Duomo 6 - 07041 Alghero (SS) - Italy

Abstract. In this paper a novel approach to identity verification, based on the analysis of face video streams, is proposed, which makes use of both physiological and behavioral features. While physical features are obtained from the subject's face appearance, behavioral features are obtained by asking the subject to vocalize a given sentence. The recorded video sequence is modelled using a Pseudo-Hierarchical Hidden Markov Model, a new type of HMM in which the emission probability of each state is represented by another HMM. The number of states are automatically determined from the data by unsupervised clustering of expressions of faces in the video. Preliminary results on real image data show the feasibility of the proposed approach.

1 Introduction

In the recent years biometrics research has grown in interest. Because of its natural interpretation (human visual recognition is mostly based on face analysis) and the low intrusiveness, face-based recognition, among others, is one of the most important biometric trait. Face analysis is a fecund research area, with a long history, but typically based on analysis of still images [15]. Recently, the analysis of video streams of face images has received an increasing attention [16, 8, 6, 3]. A first advantage in using video is the possibility of employing redundancy present in the video sequence to improve still images recognition systems, for example using voting schemes, or choosing the faces best suited for the recognition process, or also to build a 3D representation or super-resolution images. Besides these motivations, recent psychophysical and neural studies [5, 10] have shown that dynamic information is very crucial in human face recognition process. These findings inspired the development of true spatio-temporal video-based face recognition systems [16, 8, 6, 3].

All video-based approaches presented in the literature are mainly devoted to the recognition task, and to the best of our knowledge, a video-based authentication system has never been proposed. Moreover, in all video-based systems, only physiological visual cues are used: the process of recognition is based on the face appearance. When the subject is cooperative, as for authentication, also a behavioral cue can be effectively employed. For example, the subject may be

D. Zhang and A.K. Jain (Eds.): ICB 2006, LNCS 3832, pp. 113–120, 2005.

asked to vocalize a predefined sentence, such as counting from 1 to 10 or to pronounce his/her name. Each individual has its own characteristic way of vocalizing a given sentence, which could change both the appearance of the face and the temporal evolution of the visual patterns. These differences are mainly due to typical accents, pronounce, velocity of speaking, and so on. By including these behavioral features, i.e. by asking the subject to vocalize a predefined sentence, the characteristic dynamic features in the video stream are enhanced.

The system presented in this paper makes use of physiological and behavioral visual cues for person authentication, based on pseudo hierarchical Hidden Markov Models (HMM). HMMs are sequential tools largely applied in Pattern Recognition applications, and recently also employed in video-based face analysis [8, 3]. HMMs are quite appropriate for the representation of dynamic data; nonetheless, the emission probability function of a standard continuous HMM (Gaussians or Mixture of Gaussians [8, 3]) is not sufficient to fully represent the variability in the appearance of the face. In this case, it is more appropriate to apply a more complex model, such as another HMM [13, 1]. In summary, the proposed method is based on the modelling of the entire video sequence with an HMM in which the emission probability function of each state consists in another HMM itself (see Fig. 1), resulting in a pseudo-hierarchical HMM.

Determining the number of states (namely the model selection problem) is a key issue when using HMMs, and is typically selected a priori. In the method adopted, a model selection analysis has been carried out by assigning to each state of the PH-HMM a different facial expression. The problem of finding the number of states is then casted into the problem of finding all different facial expressions in the video stream. The facial expressions have been identified using an unsupervised clustering approach, where the number of clusters has been automatically determined with the Bayesian Inference Criterion [14].

2 Hidden Markov Models and Pseudo Hierarchical Hidden Markov Models

A discrete-time Hidden Markov Model $\boldsymbol{\lambda}$ can be viewed as a Markov model whose states cannot be explicitly observed: a probability distribution function is associated to each state, modelling the probability of emitting symbols from that state. More formally, a HMM is defined by the following entities [12]:

- $H = \{H_1, H_2, \cdots, H_K\}$ the finite set of the possible hidden states;
- the transition matrix $\mathbf{A} = \{a_{ij},\ 1 \leq j \leq K\}$ representing the probability to go from state H_i to state H_j;
- the emission matrix $\mathbf{B} = \{b(o|H_j)\}$, indicating the probability of the emission of the symbol o when system state is H_j (continuous or discrete)
- $\boldsymbol{\pi} = \{\pi_i\}$, the initial state probability distribution;

Given a set of sequences $\{S^k\}$, the training of the model is usually performed using the standard Baum-Welch re-estimation [12].

The evaluation step (*i.e.* the computation of the probability $P(S|\boldsymbol{\lambda})$, given a model $\boldsymbol{\lambda}$ and a sequence S to be evaluated) is performed using the *forward-backward procedure* [12].

2.1 Pseudo Hierarchical-HMM

The emission probability of a standard HMM is typically modelled using simple probability distributions, like Gaussians or Mixture of Gaussians. Nevertheless, in the case of sequences of face images, each symbol of the sequence is a face image, and a simple Gaussian could be not sufficiently accurate to properly and effectively model the probability of emission. In the PH-HMM, the emission probability is modelled using another HMM, which has been proven to be very accurate in describing faces [13, 9, 1]. The differences between standard HMMs and PH-HMM are briefly sketched in Fig. 1(a).

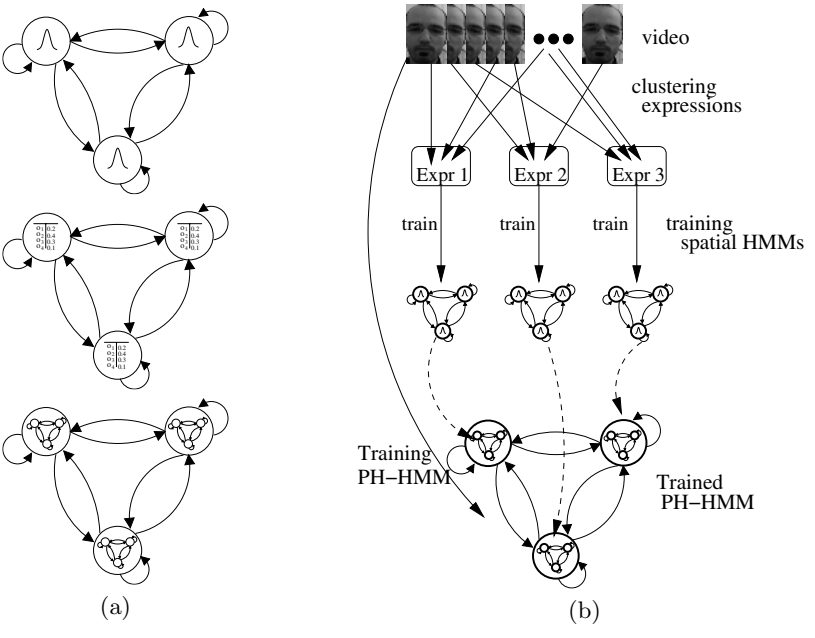

Fig. 1. (a) Differences between standard HMMs and PH-HMM, where emission probabilities are displayed into the state: (top) standard Gaussian emission; (center) standard discrete emission; (bottom) Pseudo Hierarchical HMM: in the PH-HMM the emissions are HMMs. (b) Sketch of the enrollment phase of the proposed approach.

The PH-HMM can be useful when the data have a double sequential profile. This is when the data is composed of a set of sequences of symbols $\{S^k\}$, $S^k = s_1^k, s_2^k, \cdots, s_T^k$, where each symbol s_i^k is a sequence itself: $s_i^k = o_{i1}^k, o_{i2}^k, \cdots, o_{iT_i}^k$. Let us call S^k the first-level sequences, whereas s_i^k denotes second-level sequences.

Fixed the number of states K of the PH-HMM, for each class C the training is performed in two sequential steps:

1. *Training of emission.* The first level sequence $S^k = s_1^k, s_2^k, \cdots, s_T^k$ is "unrolled", i.e. the $\{s_i^k\}$ are considered to form an unordered set U (no matter the order in which they appear in the first level sequence). This set is subsequently split in K clusters, grouping together similar $\{s_i^k\}$. For each cluster j, a standard HMM $\boldsymbol{\lambda}_j$ is trained, using the second-level sequences contained in that cluster. These HMMs $\boldsymbol{\lambda}_j$ represents the emission HMMs.
2. *Training of transition and initial states matrices.* Considering that the emission probability functions are determined by the emission HMMs, the transition and the initial states probability matrices of the PH-HMM are estimated using the first level sequences. In other words, the standard Baum Welch procedure is used, recalling that

$$b(o|H_j) = \boldsymbol{\lambda}_j$$

The number of clusters determines the number of the PH-HMM states. This value could be fixed a priori or could be directly determined from the data (using for example the Bayesian Inference Criterion [14]). In this phase, only the transition matrix and the initial state probability are estimated, since the emission has been already determined in the previous step.

Because of the sequential estimation of the PH-HMM components (firstly emission and then transition and initial state probabilities), the resulting HMM is a "pseudo" hierarchical HMM. In a truly hierarchical model, the parameters $\mathbf{A}$, $\boldsymbol{\pi}$ and $\mathbf{B}$ should be jointly estimated, because they could influence each other (see for example [2]).

3 Identity Verification from Face Sequences

Any identity verification system is based on two steps: off-line enrollment and on-line authentication.

The enrollment consists of the following sequential steps (for simplicity we assume only one video sequence $S = s_1, s_2, \cdots, s_T$, the generalization to more than one sequence is straightforward):

1. The video sequences S is analyzed to detect all faces sharing similar expression, i.e. to find clusters of expressions. Firstly, each face image s_i of the video sequence is processed, with a standard raster scan procedure, to obtain a sequence used to train a standard spatial HMM [1]. The resulting HMM models, one for each face of the video sequence, are then clustered in different groups based on their similarities [11]. Faces in the sequence with similar expression are grouped together independently from their appearance in time. The number of different expressions are automatically determined from the data using the Bayesian Inference Criterion [14].

2. For each expression cluster, a spatial face HMM is trained. In this phase all the sequences of the cluster are used to train the HMM, while in the first step one HMM for sequence has been built. At the end of the process, K HMMs are trained. We refer to these HMMs as "spatial" HMMs, because they are related to the spatial appearance of the face. In particular, each spatial HMM models a particular expression of the face in the video sequence. These models represents the emission probabilities functions of the PH-HMM.
3. The transition matrix and the initial state probability of the PH-HMM are estimated from the sequence $S = s_1, s_2, \cdots, s_T$, using the Baum-Welch procedure and the emission probabilities found in the previous step (see Sect. 2). This process aims at determining the temporal evolution of facial expressions in the video sequence. The number of states is fixed to the number of discovered clusters, this representing a sort of model selection criterion.

In summary, the main idea is to determine the facial expressions in the video sequence, modelling each of them with a spatial HMM. The expressions change during time is then modelled by the transition matrix of the PH-HMM, the "temporal" model (see Fig. 1(b))).

3.1 Spatial HMM Modelling

The process to build spatial HMMs is used in two stages of the proposed algorithm: in clustering expressions, where one HMM is trained for each face, and in the PH-HMM emission probabilities estimation, where one HMM is trained for each cluster of faces. Apart from the number of sequences used, in both cases the method consists of two steps. The former is the extraction of a sequence of sub images of fixed dimension from the original face image. This is obtained by sliding a fixed sized square window over the face image, in a raster scan fashion and keeping a constant overlap during the image scan.

For each of these sub-images, a set of low complexity features have been extracted, such as first and higher order statistics: the gray level mean, variance, Kurtosis and skewness (which are the third and the fourth moment of the data).

After the image scanning and feature extraction process, a sequence of $D \times R$ features is obtained, where D is the number of features extracted from each sub image (4), and R is the number of image patches. The learning phase is then performed using standard Baum-Welch re-estimation algorithm [12]. In this case the emission probabilities are all Gaussians, and the number of states is set to be equal to four. The learning procedure is initialized using a Gaussian clustering process, and stopped after likelihood convergence.

3.2 Clustering Facial Expressions

The goal of this step is to group together all face images in the video sequence with the same appearance, namely the same facial expression. The result is rather to label each face of the sequence corresponding to its facial expression, independently from their position in the sequence. In fact, it is possible that two

not contiguous faces share the same expression, in this sense, the sequence of faces is unrolled before the clustering process.

Since each face is described with an HMM sequence, the expression clustering process is casted into the problem of clustering sequences represented by HMMs [11, 7]. Considering the unrolled set of faces $s_1, s_2, \cdots, s_T$, where each face s_i is a sequence $s_i = o_{i1}, o_{i2}, \cdots, o_{iT_i}$, the clustering algorithm is based on the following steps:

1. Train one standard HMM $\boldsymbol{\lambda}_i$ for each sequence s_i.
2. Compute the distance matrix $D = \{D(s_i, s_j)\}$, where $D(s_i, s_j)$ is defined as:

$$D(s_i, s_j) = \frac{P(s_j|\boldsymbol{\lambda}_i) + P(s_i|\boldsymbol{\lambda}_j)}{2}$$

   This is a natural way for devising a measure of similarity between stochastic sequences. The validity of this measure in the clustering context has been already demonstrated [11].
3. Given the similarity matrix D, a pairwise distance-matrix-based method (the agglomerative complete link approach [4], in this case) is applied to perform the clustering.

In typical clustering applications the number of clusters is defined a priori. As it is impossible to arbitrarily establish the number of facial expressions in a sequence of facial images, the number of clusters has been estimated from the data, using the standard Bayesian Inference Criterion (BIC) [14], a penalized likelihood criterion.

3.3 PH-HMM Modelling

From the extracted set of facial expressions, the PH-HMM is trained. The different PH-HMM emission probability functions (spatial HMMs) model the facial expressions, while the temporal evolution of the facial expressions in the video sequence is modelled by the PH-HMM transition matrix. In particular, for each facial expression cluster, one spatial HMM is trained, using all faces belonging to the cluster (see section 3.1). The transition and the initial state matrices are estimated using the procedure described in section 2. One of the most important issues when training a HMM is model selection: in the presented approach, the number of states of the PH-HMM directly derives from the previous stage (number of clusters), representing a direct smart approach to the model selection issue.

3.4 Face Authentication

After building the PH-HMM the face authentication process, for identity verification, is straightforward. Given an unknown sequence and a claimed identity, the sequence is fed to the corresponding PH-HMM, which returns a probability value. If this value is over a predetermined threshold, the claimed identity is confirmed, otherwise it is denied.

4 Experimental Results

The system has been preliminary tested using a database composed of 5 subjects. Each subject is requested to vocalize ten digits, from one to ten. A minimum of five sequences for each subject have been acquired, in two different sessions.

The proposed approach has been tested against three other HMM-based methods, which do not fully exploit the spatio-temporal information. The first method, called "1 HMM for all", applies one spatial HMM (as described in section 3.1) to model all images in the video sequence. In the authentication phase, given an unknown video sequence, all the composing images are fed into the HMM, and the sum of their likelihoods represents the matching score. In the second method, called "1 HMM for cluster", one spatial HMM is trained for each expression cluster, using all the sequences belonging to that cluster. Given an unknown video, all images are fed into the different HMMs (and summed as before): the final matching score is the maximum among the different HMMs' scores. The last method, called "1 HMM for image", is based on training one HMM for each image in the video sequence. As in the "1 HMM for cluster" method, the matching score is computed as the maximum between the different HMMs' scores.

In all experiments only one video sequence for each subject has been used for the enrollment phase. Testing and training sets were always disjoint:in table 1 the Equal Error Rates for the four methods are reported.

Table 1. Authentication results for different methods

Method	EER
Still Image: 1 HMM for all	10.00%
Still Image: 1 HMM for cluster	11.55%
Still Image: 1 HMM for image	13.27%
Video: PH-HMM	8.275%

It is worth noting that when incorporating temporal information into the analysis a remarkable advantage is obtained, thus confirming the importance of dynamic face analysis. The applied test database is very limited and clearly too small to give a statistically reliable estimate of the performances of the method. On the other hand, the results obtained on this limited data set already show the applicability and the potential of the method in a real application scenario. The results obtained will be further verified performing a more extensive test.

5 Conclusions

In this paper a novel approach to video based face authentication is proposed, using both physiological and behavioral features. The video sequence is modelled using Pseudo Hierarchical HMM, in which the emission probability of each state

is represented by another HMM. The number of states has been determined from the data by unsupervised clustering of facial expressions in the video. The system has been preliminary tested on real image streams, showing promising results. On the other hand, more tests are required, also in comparison with other techniques, to fully evaluate the real potential of the proposed method.

References

1. M. Bicego, U. Castellani, and V. Murino. Using Hidden Markov Models and wavelets for face recognition. In *IEEE. Proc. of Int. Conf on Image Analysis and Processing*, pages 52–56, 2003.
2. S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32:41–62, 1998.
3. A. Hadid and M. Pietikäinen. An experimental investigation about the integration of facial dynamics in video-based face recognition. *Electronic Letters on Computer Vision and Image Analysis*, 5(1):1–13, 2005.
4. A.K. Jain and R. Dubes. *Algorithms for clustering data*. Prentice Hall, 1988.
5. B. Knight and A. Johnston. The role of movement in face recognition. *Visual Cognition*, 4:265–274, 1997.
6. K.C. Lee, J. Ho, M.H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2003.
7. C. Li. *A Bayesian Approach to Temporal Data Clustering using Hidden Markov Model Methodology*. PhD thesis, Vanderbilt University, 2000.
8. X. Liu and T. Chen. Video-based face recognition using adaptive hidden markov models. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2003.
9. A.V. Nefian and M.H. Hayes. Hidden Markov models for face recognition. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2721–2724, Seattle, 1998.
10. A.J. OToole, D.A. Roark, and H. Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Science*, 6:261–266, 2002.
11. A. Panuccio, M. Bicego, and V. Murino. A Hidden Markov model-based approach to sequential data clustering. In *Structural, Syntactic and Statistical Pattern Recognition*, volume LNCS 2396, pages 734–742. Springer, 2002.
12. L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of IEEE*, 77(2):257–286, 1989.
13. F. Samaria. *Face recognition using Hidden Markov Models*. PhD thesis, Engineering Department, Cambridge University, October 1994.
14. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
15. W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35:399 – 458, 2003.
16. S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91:214–245, 2003.