Available online at www.sciencedirect.com

ScienceDirect

Image and Vision Computing 27 (2009) 222–232

www.elsevier.com/locate/imavis

Dynamic face recognition: From human to machine vision

Massimo Tistarelli [a,*], Manuele Bicego [b], Enrico Grosso [b]

[a] Computer Vision Laboratory, DAP, University of Sassari, piazza Duomo 6, 07041 Alghero (SS), Italy

[b] Computer Vision Laboratory, DEIR, University of Sassari, via Torre Tonda 34, 07100 Sassari, Italy

Received 11 August 2006; received in revised form 18 March 2007; accepted 30 May 2007

Abstract

As confirmed by recent neurophysiological studies, the use of dynamic information is extremely important for humans in visual perception of biological forms and motion. Apart from the mere computation of the visual motion of the viewed objects, the motion itself conveys far more information, which helps understanding the scene. This paper provides an overview and some new insights on the use of dynamic visual information for face recognition. In this context, not only physical features emerge in the face representation, but also behavioral features should be accounted. While physical features are obtained from the subject's face appearance, behavioral features are obtained from the individual motion and articulation of the face. In order to capture both the face appearance and the face dynamics, a dynamical face model based on a combination of Hidden Markov Models is presented. The number of states (or facial expressions) are automatically determined from the data by unsupervised clustering of expressions of faces in the video. The underlying architecture closely recalls the neural patterns activated in the perception of moving faces. Experimental results obtained from real video image data show the feasibility of the proposed approach.

Keywords: Biometrics; Face recognition; Computer vision; Human perception

1. Introduction

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. On the other hand, the mutual recognition of individuals is a natural and fundamental capability of most living creatures. Therefore most living systems have a very well engineered system for the recognition of other living creatures. This is why there are many lessons we may learn from natural perceptual systems. Among them, the use of minimal energy is a natural principle of paramount importance, which is often applied in constrained optimization applied to derive numerical solutions to computer vision tasks. Many natural mechanisms also strongly rely on this principle.

This paper highlights some basic principles underlying the perceptual mechanisms of living systems, specially related to dynamic information processing, to gather insights on sensory data acquisition and processing for recognition [1].

Recently, the analysis of video streams of face images has received an increasing attention in biometric recognition [2–9]. Not surprisingly, the human visual system also implements a very sophisticated neural architecture to detect and process visual motion [30].

A first advantage in using dynamic video information is the possibility of employing redundancy present in the video sequence to improve still images recognition systems. One example is the use of voting schemes to combine results obtained for all the faces in the video, or the choice of the faces best suited for the recognition process. Another advantage is the possibility is to use the frames in a video sequence to build a 3D representation or super-resolution images. Besides these motivations, recent psychophysical and neural studies [1,10] have shown that dynamic

[*] Corresponding author. Tel.: +39 079 9720410; fax: +39 079 9720420. *E-mail address:* tista@uniss.it (M. Tistarelli).

0262-8856/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2007.05.006

information is very crucial in the human face recognition process. These findings inspired the development of true spatio-temporal video-based face recognition systems [2–9].

The case study considered in this paper makes use of both physiological and behavioral visual cues, which can be inferred from a dynamic video of human faces, for person authentication. The developed system and the theoretical framework is based on an elaboration of a classical method for sequence analysis, the *Hidden Markov Models* (HMM). The basic HMM machinery is extended to multi-dimensional data analysis in a hierarchical fashion, to build a *Pseudo-Hierarchical Hidden Markov Model* (PH-HMM). The method is based on the modeling of the entire video sequence with an HMM in which the emission probability function of each state consists in another HMM itself (see Fig. 6), resulting in a *Pseudo-Hierarchical* HMM. This complex structure represents a well founded, fully probabilistic, approach to face perception based on video modeling. This statistical tool, not only shows several interesting features related to visual recognition, but also encompasses some structural analogies with the neural architecture subduing the recognition of familiar faces in the human visual system.

Several comparative examples are presented showing the advantages of processing animated face video sequences.

2. Neurophysiology and information processing

Neural systems that mediate face recognition appear to exist very early in life. In normal infancy, the face holds particular significance and provides non-verbal information important for communication and survival [11].

The ability to recognize human faces is present during the first 6 months of life, while a visual preference for faces and the capacity for very rapid face recognition are present at birth [12,13]. By 4 months, infants recognize upright faces better than upside down faces, and at 6 months, infants show differential event-related brain potentials to familiar versus unfamiliar faces [14,15]. Apart from speech, face analysis is certainly the first and major biometric cue used by humans and therefore very important to be accurately studied.

Early studies on face recognition in primates revealed a consistent neural activity in well identified areas of the brain, mainly involving the temporal sensory area. More recent research revealed that this is not the case, but many different brain areas are taken into play at different stages of face analysis and recognition. This also recalls the need for a very complex representation including both photometric and dynamic information on the facial characteristics.

2.1. Neural mapping of face representations

Much is known about the neural systems that subserve face recognition in adult humans and primates. Face-selective neurons have been found in the inferior temporal areas (TEa and TEm), the superior temporal sensory area, the amygdala, the ventral striatum (which receives input from the amygdala) and the inferior convexity [16]. Using functional magnetic resonance imaging (fMRI), an area in the fusiform gyrus was found significantly activated when the subjects viewed faces [17–19]. Within this "general face activation area" specific regions of interest have been reported responding significantly more strongly to passive viewing of face-specific stimuli (Figs. 1 and 2). An fMRI study on individuals with autism and Asperger syndrome showed a failure to activate the fusiform face area during face processing. While a damage to fusiform gyrus and to amygdala results in impaired face recognition [20,21]. As a result, parts of the inferior and medial temporal cortex may work together to process faces. For example, the anterior inferior temporal cortex and the superior temporal sulcus project to the lateral nucleus of the amygdala, with the amygdala responsible for assigning affective significance to faces, and thus affecting both attention and mnemonic aspects of face processing [22,23].

Behavioral studies suggest that the most salient parts for face recognition are, in order of importance, eyes, mouth, and nose [24]. Eye-scanning studies in humans and monkeys show that eyes and hair/forehead are scanned more frequently than the nose [25,26], while human infants focus on the eyes rather than the mouth [27]. Using eye-tracking technology to measure visual fixations, Klin [28] recently reported that adults with autism show abnormal patterns of attention when viewing naturalistic social scenes. These patterns include reduced attention to the eyes and increased attention to mouths, bodies, and objects. The high specialization of specific brain areas for face analysis and recognition motivates the relevance of faces for social relations. On the other hand, this suggests that face understanding is not a low level process but involves higher level functional areas in the brain. These, in turn, must rely on a rich series of low level processes applied to enhance and extract face-specific features:

- *Face detection and tracking*. This process involves the analysis of dynamic as well as geometric and photometric data on the retinal projection of the face.
- *Extraction of "facial features"*. Facial features are not simply distinctive points or landmarks on the segmented face, but rather a collection of image features representing specific (and anatomically stable) areas of the face such as the eyes, eyebrows, ears, mouth, nostrils, etc. Other, subject-specific, features are also included, such as the most famous Marilyn Monroe's naevus [32].
- *Face image registration and warping*. Humans can easily recognize faces which are rotated and distorted up to a limited extent. The increase in time reported for recognition of rotated and distorted faces implies: the expectation on the geometric arrangement of facial features, and a specific process to organize the features (analogous to image registration and warping) before the actual recognition process can take place.

- *Feature matching*. This process involves the comparison between the extracted set of facial features and the same set stored in the brain. The two process of feature extraction and matching (or memory recall) are not completely separated and sequential. From the eye scan paths recorded during face recognition experiments, it seems that, after moving the eyes over few general facial features, the gaze is directed toward subject-specific features, probably to enforce the expected identity.

From these processes higher level reasoning is possible, not only to determine the subject's identity, but also to understand more abstract elements (even uncorrelated to the subject's identity) which characterize the observed person (age, race, gender, emotion, etc.). These, in turn, also require the intervention of task-specific processes, such as motion analysis and facial features tracking for understanding emotion-specific patterns [35–40].

A recent fMRI analysis on the neural architecture subduing face perception, revealed an interesting relation

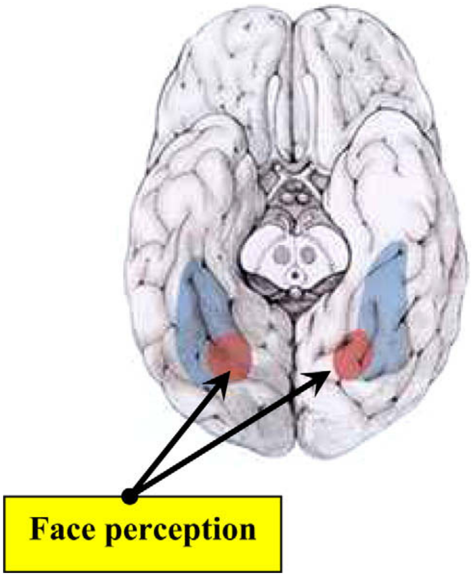

Fig. 1. Schema of the human brain as seen from below. The highlighted areas are those initially devoted to the perception of faces and object's form.

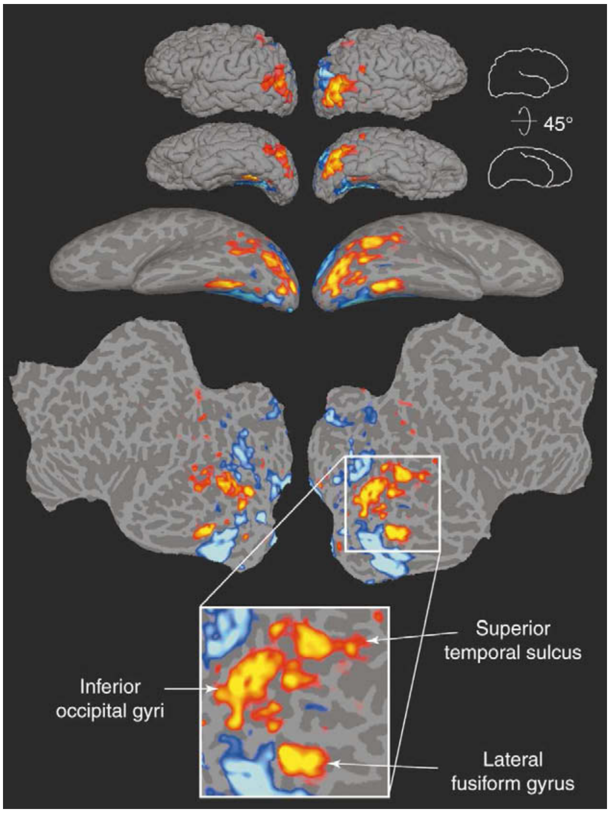

Fig. 2. Activation areas from fMRI responding mainly to face stimuli (red to yellow patterns) or to house pictures (blue patterns). (Reproduced from [29].)

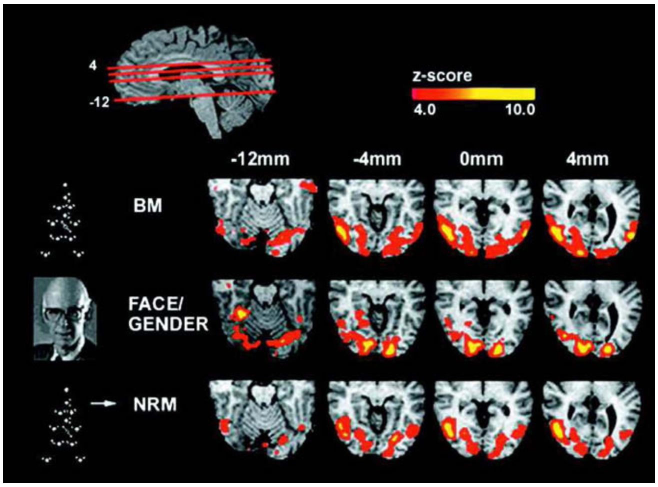

Fig. 3. Activation areas from fMRI responding to: BM biological motion (top), gender estimation from face (middle), **NRM** non-rigid motion (bottom). The red lines on the top picture indicate the position of the four axial slices spanning between −12 mm and 4 mm, with respect to the central position. On the left a schematic representation of the presented stimuli is shown: moving light dots are used for motion stimuli and face pictures for face stimuli. (Reproduced from [30].)

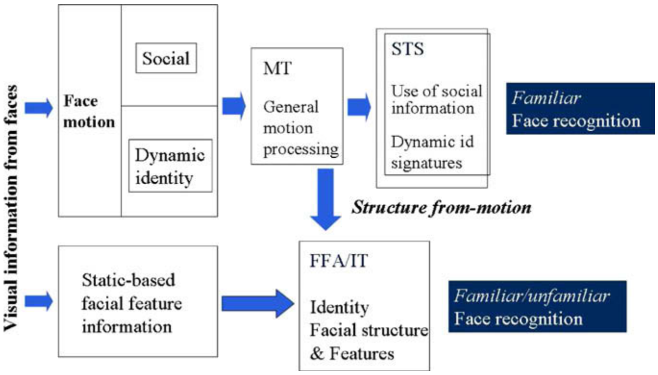

Fig. 4. Schematic representation of the perceptual processes and underlying neural systems (MT, medio temporal; IT, infero temporal; FFA, functional fusiform area, and STS, superior temporal sulcus) for dynamic analysis and recognition of familiar and unfamiliar faces (reproduced from [10]).

between perceptual tasks and neural activation. It seems that face-sensitive areas are involved also in the recognition of non-face objects such as houses, cars and animals, while specific tasks related to faces also involve non-face areas in the brain [29]. This study suggests a double architecture for face perception, formed by two connected neural activation patterns: the former devoted to process static, unchanging and invariant features of the face; the latter devoted to the analysis of changing features in the face. This architecture is also in agreement with other works on the recognition of biological motion and its relation to face perception [30], also pointing beyond the relatively narrow view suggested by earlier studies on the relation between face recognition and visual motion processing [1]. While Knight et al. discovered that motion slightly improves face recognition in difficult tasks (i.e. when 3D features are missing or the face pose is unexpected), the work by Vaina et al. also relates motion perception to face-related visual tasks. The presented fMRI study reveals a multiple activation whenever the perceptual task involves motion and shape recognition of living creatures. This also implies that the neural activation is not limited to a fixed pattern, but more strongly depends on the visual task than on the viewed subject.

As it is beyond the scope of this paper to trace all face-specific information processing, we will concentrate on the advantages of dynamic image processing for face recognition and authentication, which not only are among the most studied aspects related to visual processing human faces, but it is probably the most representative of the tasks involved in face image analysis.

2.2. Relevance of the time dimension

The high specialization of specific brain areas for face analysis and recognition motivates the relevance of faces for social relations. On the other hand, this suggests that face understanding is not a low level process but involves higher level functional areas in the brain. These, in turn, must rely on a rich series of low level processes applied to enhance and extract face-specific features. Facial features are not simply distinctive points on the segmented face, but rather a collection of image features representing specific (and anatomically stable) areas of the face such as the eyes, eyebrows, ears, mouth, nostrils, etc. Other, subject-specific, features are also included, such as the most famous Marilyn Monroe's naevus [31,32].

As shown by Vaina et al. [30], the visual task strongly influences the areas activated during visual processing. This is specially true for face perception, where not only face-specific areas are involved, but a consistent neural activity is registered in brain areas devoted to motion perception and gaze control (Fig. 3).

The time dimension is involved also when unexpected stimuli are presented [1,10,34] (Fig. 4). Humans can easily recognize faces which are rotated and distorted up to a limited extent. The increase in time reported for recognition of rotated and distorted faces implies: the expectation on the geometric arrangement of facial features, and a specific process to organize the features (analogous to image registration and warping) before the actual recognition process can take place. On the other hand, it has been shown that the recognition error for an upside-down face decreases when the face is shown in motion [1].

From the basic element related to the face shape and color, subduing a multi-area neural activity, cognitive processes are started not only to determine the subject's identity, but also to understand more abstract elements (even uncorrelated to the subject's identity) which characterize the observed person (age, race, gender, emotion, etc.). These, in turn, also recall task-specific processes, such as motion analysis and facial features tracking for understanding emotion-specific patterns [30,35–40]. As a consequence, while the motion stimuli may act as a distracter for the characterization of the identity of non-familiar faces in constrained environments. On the other hand, non-rigid and idiosyncratic facial motions constitutes a very powerful "dynamic signature" which augments the information stored for familiar faces and may indeed dramatically improve the memory recall of structured information for identity determination [33,10,34].

3. Video-based face image analysis

Conversely to previous assumptions and theories of human neural activity, face perception rarely involve a single, well defined area of the brain. It seems that the traditional "face area" is responsible for the general shape analysis but it is not sufficient for recognition as well for other tasks. In the same way, face recognition by computers can not be seen as a single, monolithic process, but several representations must be devised into a multi-layered architecture.

An interesting approach to multi-layer face processing has been proposed by Haxby [29]. The proposed architecture (sketched in Fig. 5) divides the face perception process into two main layers: the former devoted to the extraction of basic facial features and the latter processing more changeable facial features such as lip movements and expressions. It is worth noting that the encoding of changeable features of the face also captures some behavioral features of the subject, i.e. how the facial traits are changed according to a specific task or emotion.

This double-layered architecture can be represented by two distinct but similar processing units devoted to two distinct tasks. The system proposed in the remainder of the paper proposes the use of the Hidden Markov Models as elementary units to build a double layer architecture to extract shape and motion information from face sequences. The architecture is based on a multi-dimensional HMM which is capable of both capturing the shape information and the change in appearance of the face. This multi-layer architecture was termed *Pseudo Hierarchical Hidden*

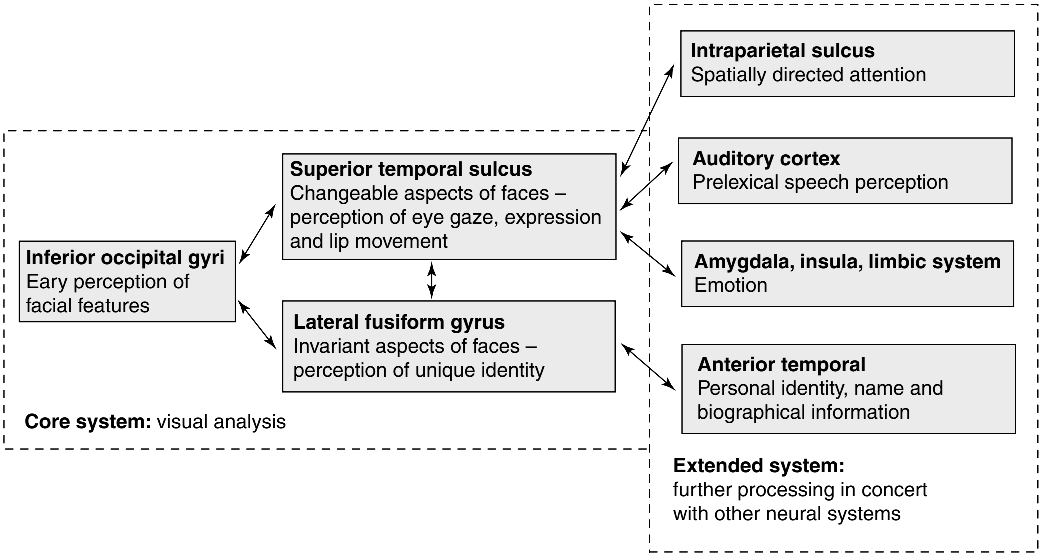
Intraparietal sulcus
Spatially directed attention
Auditory cortex
Prelexical speech perception
Superior temporal sulcus
Changeable aspects of faces – perception of eye gaze, expression and lip movement
Inferior occipital gyri
Eary perception of facial features
Amygdala, insula, limbic system
Emotion
Lateral fusiform gyrus
Invariant aspects of faces – perception of unique identity
Anterior temporal
Personal identity, name and biographical information
Core system: visual analysis
Extended system:
further processing in concert with other neural systems

Fig. 5. A model of the distributed neural system for face perception (reproduced from [29]).

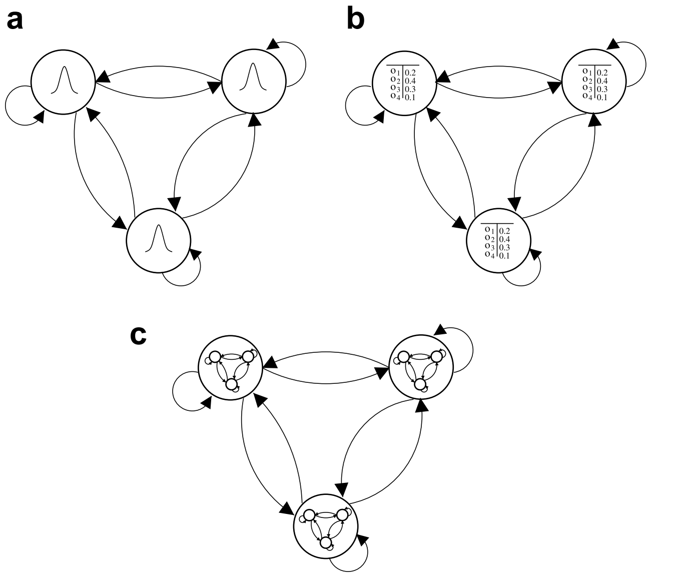

Fig. 6. Differences between standard HMMs and PH-HMM, where emission probabilities are displayed into the state: (a) standard Gaussian emission; (b) standard discrete emission; (c) Pseudo Hierarchical HMM: in the PH-HMM the emissions are HMMs.

Markov Model to emphasize the hierarchical nature of the process involved [41].

4. 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, modeling the probability of emitting symbols from that state. More formally, a HMM is defined by the following entities [42]:

- $H = \{H_1, H_2, \ldots, H_K\}$ the finite set of the possible hidden states;
- the transition matrix $\mathbf{A} = \{a_{ij}, 1 \leqslant j \leqslant K\}$ 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; typically continuous HMM were employed: $b(o|H_j)$ is represented by a Gaussian distribution;
- $\boldsymbol{\pi} = \{\pi_i\}$, the initial state probability distribution, representing probabilities of initial states;

For convenience, we denote an HMM as a triplet $\boldsymbol{\lambda} = (\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$.

Given a set of sequences $\{S^k\}$, the training of the model is usually performed using the standard Baum–Welch re-estimation [42]. During the training phase, the parameters $(\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ that maximize the probability $P(\{S^k\}|\boldsymbol{\lambda})$ are computed. 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* [42].

4.1. Pseudo Hierarchical-HMM

The emission probability of a standard HMM is typically modeled 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 may not be sufficiently accurate to properly model the probability of emission. Conversely, for the PH-HMM model, the emission probability is represented by another HMM, which has been proven to be very accurate to represent variations in the face appearance [43–46].

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 $\{S^k\}$, $S^k = s_1^k, s_2^k, \ldots, s_T^k$, where each symbol s_i^k is a sequence itself: $s_i^k = o_{i1}^k, o_{i2}^k, \ldots, o_{iT_i}^k$. 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 $S^k = s_1^k, s_2^k, \ldots, s_T^k$ 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. This process is similar to the standard Gaussian HMM initialization procedure, where the sequence is unrolled and a Mixture of K Gaussians is fitted to the unordered set. The Gaussians of the mixture are then used to roughly estimate the emission probability of each state (with a one to one correspondence with the states).

(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 [51]). 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 $\boldsymbol{A}$, $\boldsymbol{\pi}$, and $\boldsymbol{B}$ should be jointly estimated, because they could influence each other (see for example [48]).

5. Authentication of face sequences

A biometric authentication system is based on two steps: enrollment and identity verification. Given few video sequences captured from the subject's face, the enrollment phase aims at determining the best PH-HMM modeling the subject's face appearance. This model encompasses both the invariant aspects of the face and its changeable features. Identity verification is performed by projecting a captured face video sequence on the PH-HMM model belonging to the claimed identity.

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

(1) The video sequence 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 reduced to a raster scan sequence of pixels, used to train a standard spatial HMM [43,46]. The resulting face HMM models are clustered in different groups based on their similarities [49,50]. 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 [51].

(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. At the end of the process, K HMMs are trained. 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 $S = s_1, s_2, \ldots, s_T$, using the Baum–Welch procedure and the emission probabilities found in the previous step (see Section 4). This process aims at determining the temporal evolution of facial expressions over time. The number of states is fixed to the number of discovered clusters, this representing a sort of model selection criterion.

In summary, the main objective of the PH-HMM representation scheme is to determine the facial expressions in the video sequence, modeling each of them with a spatial HMM. The expressions change during time is then modeled by the transition matrix of the PH-HMM, which constitutes the "temporal" model (as sketched in Fig. 7).

5.1. Spatial HMM modeling: analysis of face form

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. Independently of the number of sequences used, in both cases the method involves 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 (Fig. 8).

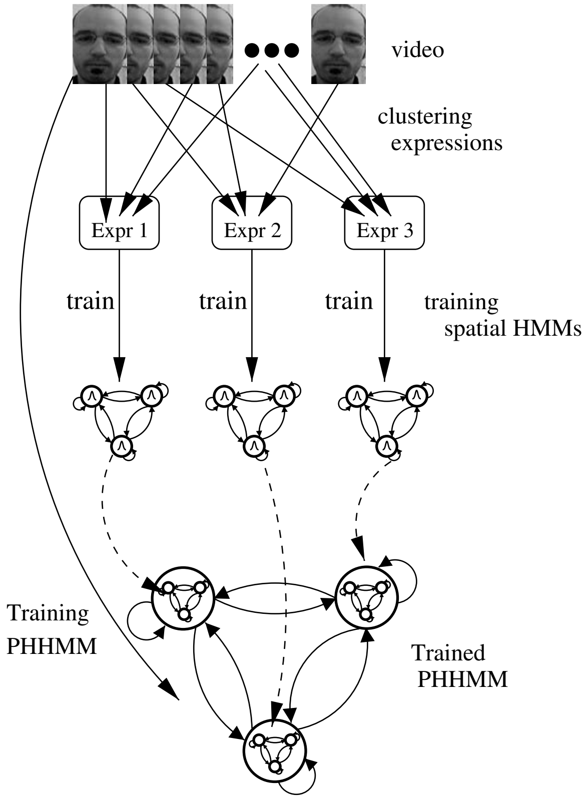

Fig. 7. Sketch of the enrollment phase of the proposed approach.

This scheme is similar to the one proposed in [43,46] for recognition. The main difference stems from the extracted features. In [43,46] the Discrete Cosine Transform (DCT) and wavelets were applied to obtain the facial features. These features, which demonstrated to be very robust for classification, are computationally very demanding, making the method not easily applicable in real operational environments. In the present approach, the local image structure is captured computing first and higher order statistics: the gray level mean, variance, Kurtosis and skewness (which are the third and the fourth moment of the data) [47].

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 (four features in total), and R is the number of image patches. The learning phase is then performed using standard Baum–Welch re-estimation algorithm [42]. 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.

5.2. 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. It is worth noting that this process does not imply a segmentation of the sequence into homogeneous, contiguous fragments. 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, as for example pronouncing the two “w” in the word “twentytwo”. 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 [52,49,53,50]. Considering the unrolled set of faces $s_1, s_2, \ldots, s_T$, where each face s_i is a sequence $s_i = o_{i1}, o_{i2}, \ldots, o_{iT_i}$, 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. Since $\boldsymbol{\lambda}_i$ is trained using the sequence s_i, the closer is s_j to s_i, the higher is the probability $P(s_j|\boldsymbol{\lambda}_i)$. Please note that this is not a quantitative but rather a qualitative measure of similarity. The validity of this measure in the clustering context has been already demonstrated [50,49].

(3) Given the similarity matrix D, a pairwise distance-matrix-based method (e.g. an agglomerative method) is applied to perform the clustering. In particular, the agglomerative complete link approach [54] has been used.

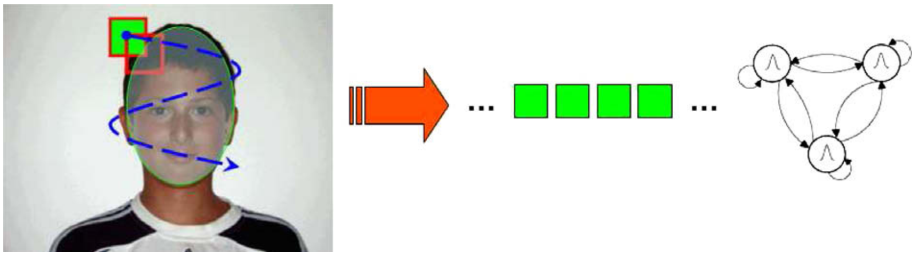

Fig. 8. Sampling scheme applied to generate the sequence of sub-images and the HMM model of the sampled sequence, representing a single face image.

In typical clustering applications the number of clusters is defined a priori. In this application, it is practically impossible (or not viable in many real cases) to arbitrarily establish the number of facial expressions which may appear in a sequence of facial images. Therefore, the number of clusters has been estimated from the data, using the standard Bayesian Inference Criterion (BIC) [51]. This is a penalized likelihood criterion which is able to find the best number of clusters as the compromise between the model fitting (HMM likelihood) and the model complexity (number of parameters).

5.3. PH-HMM modeling: analysis of temporal evolution

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 modeled 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 5.1). The transition and the initial state matrices are estimated using the procedure described in Section 4.

One of the most important issues when training a HMM is the model selection, or the estimation of the best number of states. In fact, this operation can prevent overtraining and undertraining which may lead to an incorrect model representation. 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.

5.4. Face verification

The verification of a subject's identity is straightforward. Captured a sequence of face images from an unknown subject, and a claimed identity, the sequence is fed to the corresponding PH-HMM, which returns a probability value. The claimed identity is verified if the computed probability value is over a predetermined threshold. This comparison corresponds to verifying if the captured face sequence is well modeled by the given PH-HMM.

6. Experimental results

The system has been tested using a database composed of 21 subjects. During the video acquisition, each subject was requested to vocalize ten digits, from one to ten. A minimum of five sequences for each subject have been acquired, in two different sessions. Each sampled video is composed of 95–195 color images, with several changes in facial expression and scale (see Fig. 9). The images have a resolution of 640×480 pixels. For the face classification experiments the images have been reduced to gray level with 8 bits per pixel.

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 5.1) 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

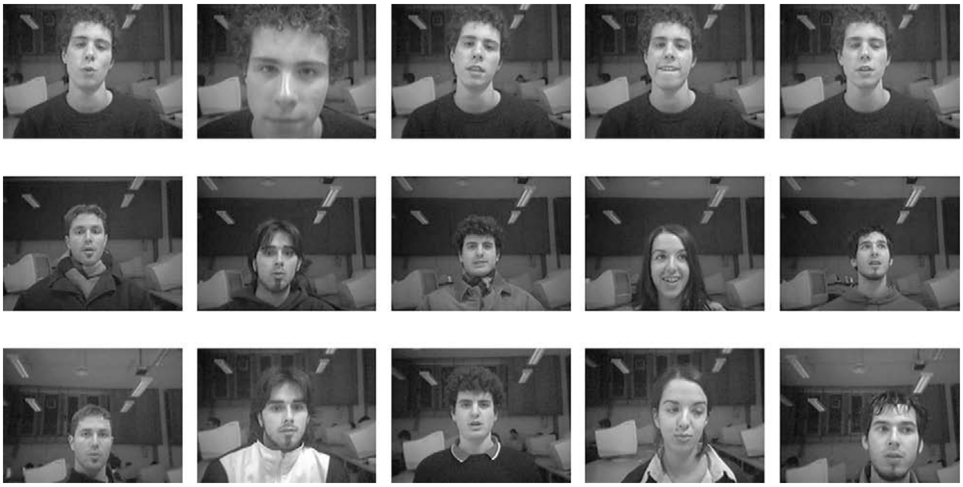

Fig. 9. (Top) Example frames of one subject extracted from the collected video database. (Middle) One sample frame of five subjects, extracted from the first acquisition session. (Bottom) One sample frame of the same subjects above, extracted from the second acquisition session.

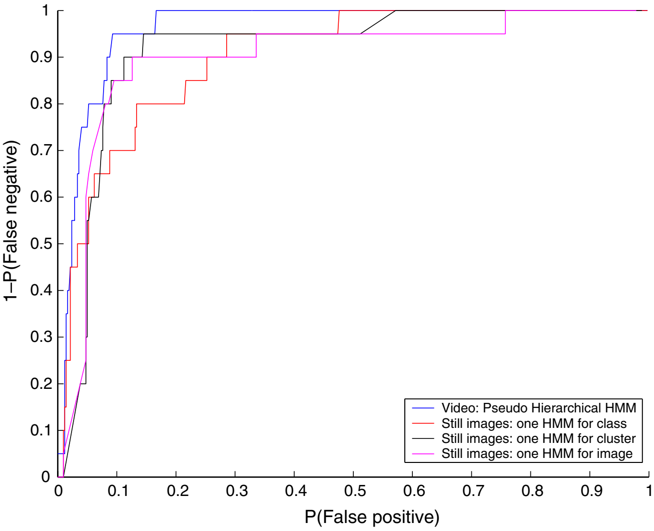

Fig. 10. The computed ROC curve for the verification experiment from video sequences of faces for the four methods reported.

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. Full client and impostor tests have been performed computing a Receiving Operating Characteristic (ROC) curve (Fig. 10). Testing and training sets were always disjoint, allowing a more reliable estimation of the error rate. In Table 1 the Equal Error Rates (error when false positive and false negatives are equal) for the four methods are reported.

The analysis of the video sequences with the hierarchical, spatio-temporal HMM model produced a variable number of clusters, varying from 2 to 10, depending on the coding produced by the spatial HMMs. It is worth noting that when incorporating temporal information into the analysis a remarkable advantage is obtained, thus confirming the importance of explicitly modeling the face motion for identification and authentication (Tables 1 and 2).

Table 1
Verification results for the reported HMM based, face modeling methods

Method	EER (%)
Still image: 1 HMM for all	20.24
Still image: 1 HMM for cluster	10.60
Still image: 1 HMM for image	13.81
Video: PH-HMM	6.07

Table 2
Identification results (accuracy) for the reported HMM based, face modeling methods

Method	Accuracy
Still image: 1 HMM for all	52.38% (11/21)
Still image: 1 HMM for cluster	66.67% (14/21)
Still image: 1 HMM for image	57.14% (12/21)
Video: PH-HMM	100% (21/21)

Table 3
Identification (accuracy) and verification (EER) results for the reported baseline PCA-based recognition methods

Combination rule	Accuracy	EER (%)
PCA with MAX rule	66.67% (14/21)	9.17
PCA with SUM rule	95.23% (20/21)	5

6.1. Baseline testing

In order to compare the performances of the PH-HMM method with standard recognition algorithms, the same image sequences have been analyzed with a baseline Principal Component Analysis (PCA) algorithm. In order to integrate the information from all images in the sequence, the MAX rule and the SUM rule were applied to combine the scores from single images. The results obtained are reported in Table 3.

As it can be noted, while the PCA with the MAX rule provides very poor results (comparable to modeling the sequence with one HMM per cluster), the SUM rule provides results which are very similar to the PH-HMM. These results confirm the validity of the proposed recognition scheme which demonstrated its capability to capture both shape and dynamic information on the face.

The applied test database is very limited and clearly too small to give a statistically reliable estimate of the performances of the method. Nonetheless, the results obtained on this limited data set already show the applicability and the potential of the method in a real application scenario. On the other hand, the tests performed on this limited dataset allowed to compare different modeling schemes where the face dynamics was loosely integrated into the computational model. The proposed PH-HMM model outperforms all other modeling schemes based on the HMMs, at the same time it represents a very interesting computational implementation of the human model of face recognition, as proposed by Haxby in [29] and described in Section 3. It is important to stress that, far from being the best computational solution for face recognition of faces from video, the proposed scheme closely resembles the computational processes underlying the recognition of faces in the human visual system.

In order to further investigate the real potential of the proposed modeling scheme, the results obtained will be further verified performing a more extensive test on a database including at least 50 subjects and 10 image sequences for each subject.

7. Conclusions

Despite of the simple neural architectures for face perception hypothesized in early neurological studies, the perception of human faces is a very complex task which involves several areas of the brain. The neural activation pattern depends on the specific task required rather than on the nature of the stimulus. This task-driven model may be represented by a dual layer architecture where static and dynamic features are analyzed separately to devise a unique face model.

The dual nature of the neural architecture, subduing face perception, allows to capture both static and dynamic data. As a consequence, not only physiological features are processed, but also behavioral features, which are related to the way the face traits are changing over time. This last property is characteristic of each individual and implicitly represents the changeable features of the face.

A statistical model of the face appearance, which reflects the described dual-layered neural architecture, has been presented. In order to capture both static and dynamic features, the model is based on the analysis of face video sequences using a multi-dimensional extension of Hidden Markov Models, called Pseudo Hierarchical HMM. In the PH-HMM model, the emission probability of each state is represented by another HMM, while the number of states is determined from the data by unsupervised clustering of facial expressions in the video. The resulting architecture is then capable of modeling both physiological and behavioral features, represented in the face image sequence and well represents the dual neural architecture described by Haxby in [29]. It is worth noting that the proposed approach far from being the best performing computational solution for face recognition from video, has been explicitly devised to copy the neural processes subduing face recognition in the human visual system.

Even though the experiments performed are very preliminary, already demonstrate the potential of the algorithm in coupling photometric appearance of the face and the temporal evolution of facial expressions. The proposed approach can be very effective in face identification or verification to exploit the subject's cooperation in order to enforce the required behavioral features and strengthen the discrimination power of a biometric system.

References

[1] B. Knight, A. Johnston, The role of movement in face recognition, Visual Cognition 4 (1997) 265–274.

[2] O. Yamaguchi, K. Fukui, K. Maeda, Face recognition using temporal image sequence, in: Proc. Int. Conf. Automatic Face and Gesture Recognition, 1998.

[3] Z. Biuk, S. Loncaric, Face recognition from multi-pose image sequence, in: Proc. Int. Symp. Image and Signal Processing and Analysis, 2001.

[4] Y. Li, Dynamic face models: construction and applications. Ph.D. thesis, Queen Mary, University of London, 2001.

[5] G. Shakhnarovich, J.W. Fisher, T. Darrell, Face recognition from long-term observations, in: Proc. European Conf. Computer Vision, 2002.

[6] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Computer Vision and Image Understanding 91 (2003) 214–245.

[7] X. Liu, T. Chen, Video-based face recognition using adaptive hidden markov models, in: Proc. Int. Conf. Computer Vision and Pattern Recognition, 2003.

[8] K.C. Lee, J. Ho, M.H. Yang, D. Kriegman. Video-based face recognition using probabilistic appearance manifolds, in: Proc. Int. Conf. Computer Vision and Pattern Recognition, 2003.

[9] A. Hadid, 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) (2005) 1–13.

[10] A.J. OToole, D.A. Roark, H. Abdi, Recognizing moving faces: a psychological and neural synthesis, Trends in Cognitive Science 6 (2002) 261–266.

[11] C. Darwin, The Expression of the Emotions in Man and Animals, John Murray, London, UK, 1965, Original work published 1872.

[12] C. Goren, M. Sarty, P. Wu, Visual following and pattern discrimination of face-like stimuli by newborn infants, Pediatrics 56 (1975) 544–549.

[13] G.E. Walton, T.G.R. Bower, Newborns form prototypes in less than 1 minute, Psychological Science 4 (1993) 203–205.

[14] J. Fagan, Infants recognition memory for face, Journal of Experimental Child Psychology 14 (1972) 453–476.

[15] M. de Haan, C.A. Nelson, Recognition of the mothers face by 6-month-old infants: a neurobehavioral study, Child Development 68 (1997) 187–210.

[16] C.M. Leonard, E.T. Rolls, F.A.W. Wilson, G.C. Baylis, Neurons in the amygdala of the monkey with responses selective for faces, Behavioral Brain Research 15 (1985) 159–176.

[17] I. Gauthier, M.J. Tarr, A.W. Anderson, P. Skudlarski, J.C. Gore, Activation of the middle fusiform face area increases with expertise in recognizing novel objects, Nature Neuroscience 2 (1999) 568–573.

[18] N. Kanwisher, J. McDermott, M.M. Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception, Journal of Neuroscience 17 (1997) 4302–4311.

[19] G. McCarthy, A. Puce, J.C. Gore, T. Allison, Face specific processing in the human fusiform gyrus, Journal of Cognitive Neuroscience 8 (1997) 605–610.

[20] R.T. Schultz, I. Gauthier, A. Klin, R.K. Fulbright, A.W. Anderson, F.R. Volkmar, P. Skudlarski, C. Lacadie, D.J. Cohen, J.C. Gore, Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and Asperger syndrome, Archives of General Psychiatry 57 (2000) 331–340.

[21] A.R. Damasio, J. Damasio, G.W. Van Hoesen, Prosopagnosia: anatomic basis and behavioral mechanisms, Neurology 32 (1982) 331–341.

[22] C.A. Nelson, The development and neural bases of face recognition, Infant and Child Development 10 (2001) 3–18.

[23] J.P. Aggleton, M.J. Burton, R.E. Passingham, Cortical and subcortical afferents to the amygdala of the rhesus monkey (*Macaca mulatta*), Brain Research 190 (1980) 347–368.

[24] J. Shepherd, Social factors in face recognition, in: G. Davies, H. Ellis, J. Shepherd (Eds.), Perceiving and Remembering Face, Academic Press, London, 1981, pp. 55–79.

[25] L. Yarbus, Eye Movements and Vision, Plenum Press, New York, 1967.

[26] F.K.D. Nahm, A. Perret, D. Amaral, T.D. Albright, How do monkeys look at faces? Journal of Cognitive Neuroscience 9 (1997) 611–623.

[27] M.M. Haith, T. Bergman, M.J. Moore, Eye contact and face scanning in early infancy, Science 198 (1979) 853–854.

[28] A. Klin, Eye-tracking of social stimuli in adults with autism, in: NICHD Collaborative Program of Excellence in Autism, Yale University, New Haven, CT, 2001.

[29] J.V. Haxby, E.A. Hoffman, M.I. Gobbini, The distributed human neural system for face perception, Trends in Cognitive Sciences 4 (6) (2000) 223–233.

[30] L.M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, J.W. Belliveau, Functional neuroanatomy of biological motion perception in humans, in: Proc. National Academy of Sciences of the United

States of America, vol. 98, No. 20, Sep. 25, 2001, pp. 11656–11661.

[31] M. Tistarelli, E. Grosso, Active vision-based face authentication, in: M. Tistarelli (Ed.), Image and Vision Computing: Special Issue on Facial Image Analysis, vol. 18, No. 4, 2000, pp. 299–314.

[32] M. Bicego, E. Grosso, M. Tistarelli, On finding differences between faces, in: T. Kanade, A. Jain, N.K. Ratha (Eds.), Audio- and Video-based Biometric Person Authentication, LNCS, vol. 3546, Springer, 2005, pp. 329–338.

[33] A. Pike, Face Perception in Motion, Personal Communication, 2006.

[34] A. O'Toole, Face Perception in Humans and Machines, Personal Communication (2006).

[35] L. Wiskott, J.M. Fellous, N. Kruger, C. von der Malsburg, Face recognition and gender determination, in: Proc. Int. Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, 1995, pp. 92–97.

[36] H. Wechsler, P. Phillips, V. Bruce, F. Soulie, T. Huang (Eds.), Face Recognition from Theory to Applications, NATO ASI Series F, vol. 163, Springer-Verlag, Berlin, Heidelberg, 1998.

[37] G. Cottrell, J. Metcalfe, Face gender and emotion recognition using holons, in: D. Touretzky (Ed.), Advances in Neural Information Processing Systems, vol. 3, Morgan Kaufmann, San Mateo, CA, 1991, pp. 564–571.

[38] B. Braathen, M.S. Bartlett, G. Littlewort, J.R. Movellan, First Steps Towards Automatic Recognition of Spontaneous Facial Action Units ACM Workshop on Perceptive User Interfaces, Orlando, FL, Nov. 15–16 2001.

[39] R.W. Picard, Toward computers that recognize and respond to user emotion, IBM System 3/4 (39) (2000).

[40] R.W. Picard, Building HAL: Computers that sense, recognize, and respond to human emotion, MIT Media-Lab TR-532, Also in Society of Photo-Optical Instrumentation Engineers, Human Vision and Electronic Imaging VI, part of SPIE9s Photonics West, 2001.

[41] M. Bicego, E. Grosso, M. Tistarelli, Person authentication from video of faces: A behavioral and physiological approach using Pseudo Hierarchical Hidden Markov Models, in: Proc. Int. Conf. Biometric Authentication 2006, LNCS 3832, Springer-Verlag, Hong Kong, China, 2006, pp. 113–120.

[42] L. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of IEEE 77 (2) (1989) 257–286.

[43] V.V. Kohir, U.B. Desai, Face recognition using DCT-HMM approach, in: Proc. Workshop on Advances in Facial Image Analysis and Recognition Technology (AFIART), Freiburg, Germany, 1998.

[44] F. Samaria, Face recognition using Hidden Markov Models, Ph.D. thesis, Engineering Department, Cambridge University, October 1994.

[45] A.V. Nefian, M.H. Hayes, Hidden Markov models for face recognition, in: Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Seattle, 1998, pp. 2721–2724.

[46] M. Bicego, U. Castellani, V. Murino, Using Hidden Markov Models and wavelets for face recognition, in: IEEE. Proc. Int. Conf. Image Analysis and Processing, 2003, pp. 52–56.

[47] M. Bicego, E. Grosso, M. Tistarelli, Probabilistic face authentication using hidden markov models, in: Proc. SPIE Int. Workshop on Biometric Technology for Human Identification, 2005.

[48] S. Fine, Y. Singer, N. Tishby, The hierarchical hidden markov model: analysis and applications, Machine Learning 32 (1998) 41–62.

[49] P. Smyth, Clustering sequences with hidden Markov models, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, MIT Press, 1997, p. 648.

[50] A. Panuccio, M. Bicego, V. Murino, A Hidden Markov model-based approach to sequential data clustering, in: Structural, Syntactic and Statistical Pattern Recognition, LNCS 2396, Springer, 2002, pp. 734–742.

[51] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics 6 (2) (1978) 461–464.

[52] L. Rabiner, C. Lee, B. Juang, J. Wilpon, HMM clustering for connected word recognition, in: Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 1989, pp. 405–408.

[53] C. Li, A Bayesian Approach to Temporal Data Clustering using Hidden Markov Model Methodology. Ph.D. thesis, Vanderbilt University, 2000.

[54] A.K. Jain, R. Dubes, Algorithms for Clustering Data, Prentice Hall, 1988.