Recognition of Human Faces: From Biological to Artificial Vision

Massimo Tistarelli[1], Linda Brodo[2], Andrea Lagorio[3], and Manuele Bicego[3]

[1] DAP - University of Sassari, piazza Duomo 6 - 07041 Alghero (SS) - Italy
tista@uniss.it

[2] DSL - University of Sassari, piazza Università 21 - 07100 Sassari - Italy
brodo@uniss.it

[3] DEIR - University of Sassari, via Torre Tonda 34 - 07100 Sassari - Italy
{lagorio,bicego}@uniss.it

Abstract. Face recognition is among the most challenging techniques for personal identity verification. Even though it is so natural for humans, there are still many hidden mechanisms which are still to be discovered. According to the most recent neurophysiological studies, the use of dynamic information is extremely important for humans in visual perception of biological forms and motion. Moreover, motion processing is also involved in the selection of the most informative areas of the face and consequently directing the attention. This paper provides an overview and some new insights on the use of dynamic visual information for face recognition, both for exploiting the temporal information and to define the most relevant areas to be analyzed on the face. In this context, both physical and behavioral features emerge in the face representation.

1 Introduction

Biometric recognition has attracted the attention of scientists, investors, government agencies as well as the media for the great potential in many application domains. It turns out that there are still a number of intrinsic drawbacks in all biometric techniques. In this talk we postulate the need for a proper data representation which may simplify and augment the discrimination among different instances or biometric samples of different subjects. In fact, considering the design of many natural systems, it turns out that spiral (circular) topologies are the best suited to economically store and process data. Among the many developed techniques for biometric recognition, face analysis seems to be the most promising and interesting modality. The ability of the human visual system of analyzing unknown faces, is an example of the amount of information which can be extracted from face images. This is not limited to the space or spectral domain, but heavily involves the time evolution of the visual signal. Nonetheless, there are still many open problems which need to be faced as well. This not only requires to devise new algorithms but to determine the real potential and limitations of existing techniques, also exploiting the time dimensionality to boost recognition performances.

F. Mele et al. (Eds.): BVAI 2007, LNCS 4729, pp. 191–213, 2007.

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 [2,3,4,5,6,7,8,9]. Not surprisingly, the human visual system also implements a very sophisticated neural architecture to detect and process visual motion [10].

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 [1,11] have shown that dynamic information is very crucial in the human face recognition process. These findings inspired the development of true spatio-temporal video-based face recognition systems [2,3,4,5,6,7,8,9]. Last, but not least, the recognition of faces in the human visual system also involves attention mechanisms to detect and analyze the “most salient” features in the face. How these features are defined and detected is still not completely understood. Nonetheless, very distinctive information are used to characterize human faces. A computer implementation is introduced where salient regions are defined by analyzing several individuals. A set of multi-scale patches are extracted from each face image before projecting them into a common feature space. The degree of “distinctiveness” of any patch depends on its distance in feature space from patches mapped from other individuals. Both a perceptual experiment, involving 45 observers and a technological experiment were performed and compared. A further comparative analysis showed that the performance of the n-ary approach is as good as several contemporary unary, or binary, methods - whilst tapping a complementary source of information.

2 Human Vision 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 nonverbal information important for communication and survival [12].

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 [13,14]. 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 [15,16]. 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

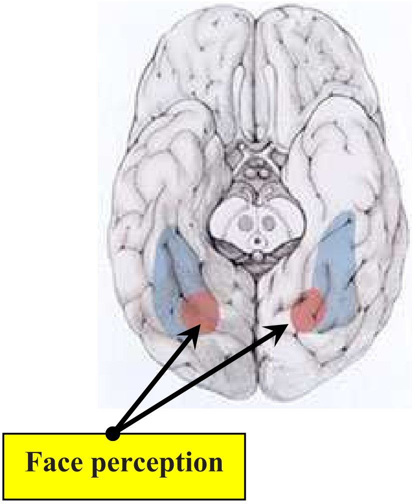


Fig. 1. Picture of the human brain as seen from below. The highlighted areas are those initially devoted to the perception of faces and object's form.

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 Space-Variant Image Representations

To achieve any visual task, including face recognition, humans are able to purposively control the flow of input data limiting the amount of information gathered from the sensory system [17,18,19]. This is needed to reduce the space and computation time required to process the incoming information. The anatomy of the early stages of the human visual system is a clear example: despite the formidable acuity in the fovea centralis (1 minute of arc) and the wide field of view (about 140x200 degrees of solid angle), the optic nerve is composed of only 10^6 nerve fibres. The space-variant distribution of the ganglion cells in the retina allows a formidable data flow reduction. In fact, the same resolution would result in a space-invariant sensor of about $6x10^8$ pixels, thus resulting in a compression ratio of 1:600 [20]. The probability density of the spatial distribution of the ganglion cells, which convey the signal from the retinal layers to the optic nerve and is responsible for the data compression, follows a logarithmic-polar law. The number of cells decreases from the center of the retina toward the periphery, with the maximal resolution in the fovea [21]. The same data compression can be obtained on electronic images, either by using a specially designed space-variant sensor [22], or re-sampling a standard image according to the log-polar transform [19,20]. The analytical formulation of the log-polar mapping describes

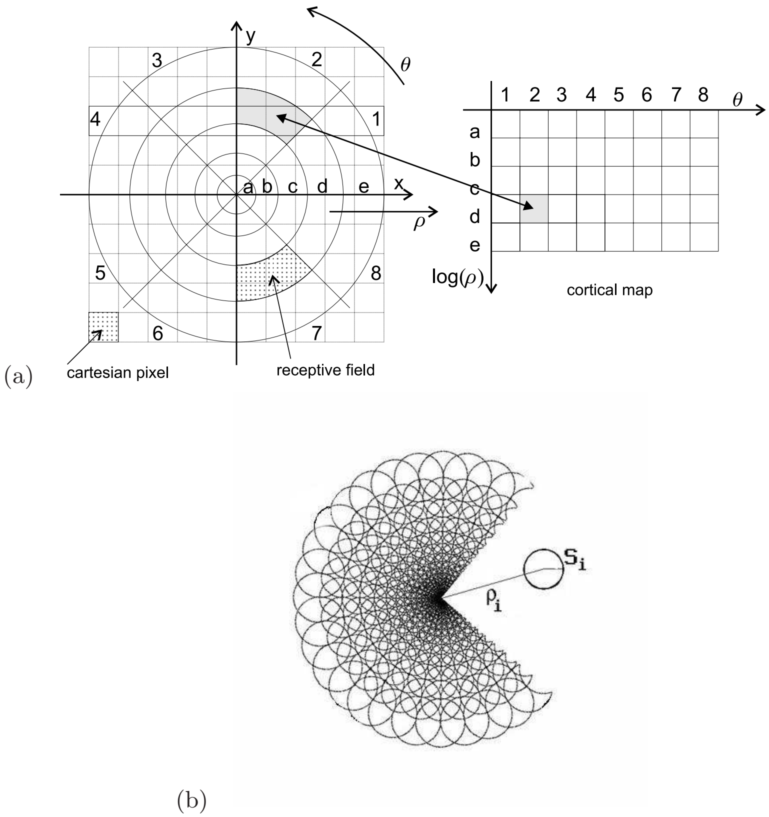


Fig. 2. (a) Log-polar sampling for Cartesian image remapping and (b) discrete log-polar model

the mapping that occurs between the retina (retinal plane (ρ, θ)) and the visual cortex (log-polar or cortical plane (ξ, η)). The derived logarithmic-polar law, taking into account the linear increment in size of the receptive fields, from the central region (fovea) towards the periphery, is given by:

$$\begin{cases} x = \rho \cos \theta \\ y = \rho \sin \theta \end{cases} \qquad\qquad \begin{cases} \eta = q\ \theta \\ \xi = \ln_a \frac{\rho}{\rho_0} \end{cases} \tag{1}$$

where a defines the amount of overlap among neighboring receptive fields, ρ_0 is the radius of the innermost circle, $\frac{1}{q}$ is the minimum angular resolution of the log-polar layout, and (ρ, θ) are the polar coordinates of an image point.

Other models for space-variant image geometries have been proposed, like the truncated pyramid [23], the reciprocal wedge transform (RWT) [24] and the complex logarithmic mapping (CLM) [25]. Several implementations of space-variant imaging have been developed: space-variant sensors [22], custom designed image re-sampling hardware [26], and special software routines [19,27]. Given the high processing power of current computing hardware, image re-mapping can be performed at frame rate without the need of special computing hardware, and also allows the use of conventional, low cost, cameras.

3 Visual Attention and Selective Processing

A very general and yet very important perceptual mechanism in humans is visual attention [28]. This mechanism is exploited by the human perceptual system to parse the input signal in various dimensions: "signal space" (low or high frequency data), depth (image areas corresponding to objects close or far from the observer), motion (static or moving objects) etc. The selection is controlled through ad-hoc band-limiting or focusing processes, which determine the areas of interest in the scene to which direct the gaze [29].

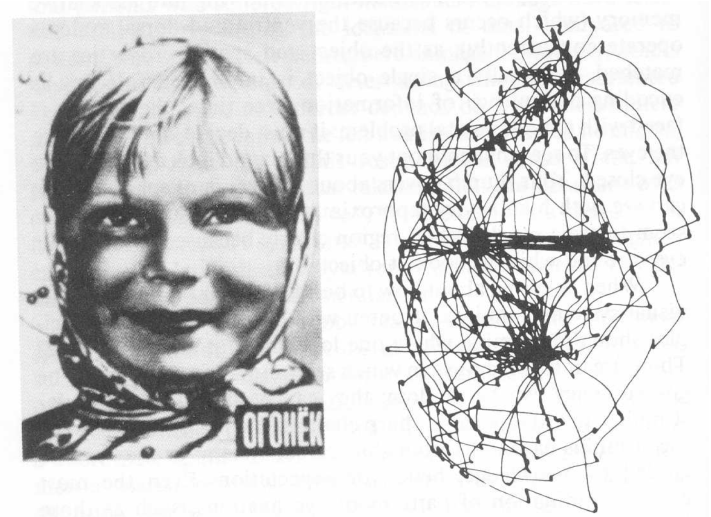


Fig. 3. Schema of the saccades performed by the human visual system analyzing an unfamiliar face (reprinted from [28])

In the case of face perception, both space-variant image re-sampling and the adoption of a selective attention mechanism can greatly improve the performance of any recognition/authentication algorithm. While the log-polar mapping allows to adaptively reduce the frequency content of the input signal, more sophisticated processes are needed to discard low information areas in the image. Visual attention in humans is also devoted to detect the most informative areas in the face to produce a compact representation for higher level cognitive processes.

Behavioral studies suggest that, in general, the most salient parts for face recognition are, in order of importance, eyes, mouth, and nose [30]. Eye-scanning studies in humans and monkeys show that eyes and hair/forehead are scanned more frequently than the nose [28,31], while human infants focus on the eyes rather than the mouth [32]. Using eye-tracking technology to measure visual fixations, Klin [33] 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 further demonstrates that face understanding is not a low level process but involves higher level functional areas in the brain.

Even though visual attention is generally focused on almost fixed facial landmarks, this does not imply that these are the only areas processed for face perception. 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. Two different kind of landmarks can be defined:

- face-invariant landmarks, such as the eyes, the nose, the mouth, the ears and all other elements which are typical of every face;
- face-variant landmarks, which are distinctive elements for a given subject's face [34,35].

The face-invariant landmarks are important to distinguish faces from non-faces, and constitute the basic elements to describe both familiar and unfamiliar faces. All face-variant landmarks constitute the added information, which is learned by the human visual system, to uniquely characterize a subject's face. As a consequence, attention is selectively driven to different areas of the face corresponding to the subject's specific landmarks. This hypothesis is grounded, not only on considerations related to the required information processing, but also on several observations of the eye movements while processing human faces [13,28,31,32,33]. In all reported tests, the gaze scanpaths were different according to the identity of the presented face. As a consequence, the classification of subjects based on the face appearance, must be tuned to extract and process the most salient features of the face itself.

3.1 A Computational Model for Selective Face Processing

In order to define distinctive or salient areas of an individual's face a comparative analysis is made. All the areas of an individual's face, that appear distinct when compared to other faces from the population, are selected.

Because the appearance of different subjects is compared, this approach is conceptually different from most of the existing feature extraction methods that rely on the detection and analysis of specific face areas for authentication or recognition purposes—e.g. the Elastic Bunch Graph Matching technique [36]. It differs also from more elaborate techniques that identify the most "salient" parts within the face according to a pre-specified criterion. Among these [37,38,39,40], the system described by [41] that detects "key points" from a set of lines extracted from the face image and that in [42] which selects "characteristic points"

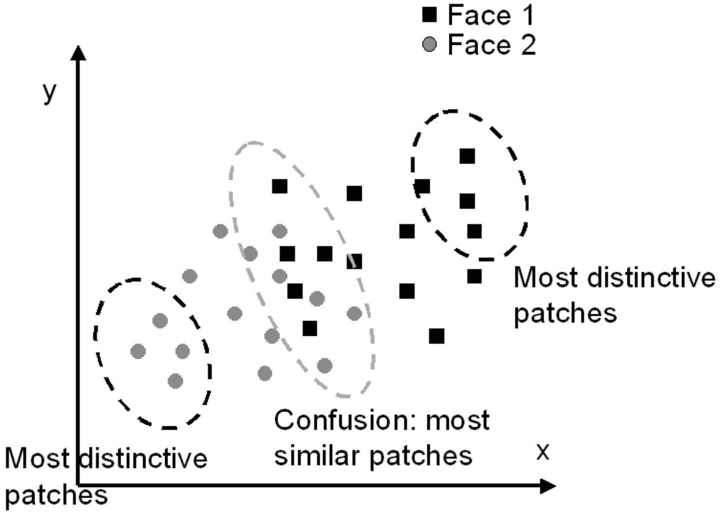


Fig. 4. Schema describing the pair-wise differences algorithm. The x and y axes represent two hypothetical coordinates in the feature space.

in a generic image by means of a local optimization process applied to the difference of Gaussians image, filtered at different scales and orientations. Though they all vary in implementation, robustness, computational requirements and accuracy, each of the above approaches is essentially a *unary* technique: salient regions are defined by analyzing *only one* instance of the face class, namely only images of the *same* individual. On the contrary, we identify local patches within an individual's face that are *different* from other individuals by performing a pair-wise, or binary, analysis. This avoids issues that may arise when invoking a single average face, or canonical model, against which each face would then be distinguished. In particular, differences between faces are determined by directly extracting from one individual's face image the most distinguishing or dissimilar patches with respect to another's. Image patches from the same individual tend to cluster together when projected in a multi-dimensional space and the distance, in that space, of that patch from clusters formed by other faces can be used as a measure of "distinctiveness"—as sketched, in just 2-D, in Fig. 8.

It is worth noting that the concept of comparative face analysis is also inherent in the work by Penev and Atik [43] (Local Feature Analysis), as well as by Li *et al.* [44] (Local Nonnegative Matrix Factorization), and by Kim *et al.* [45] (Locally Salient Independent Component Analysis). These are locally salient versions of dimensionality reduction techniques, applied to a database of images so to obtain a local representation (as a set of basis) of the training set. Even if not explicitly developed to extract salient parts of a face, all these techniques find utility in characterizing a face by performing a comparative local analysis.

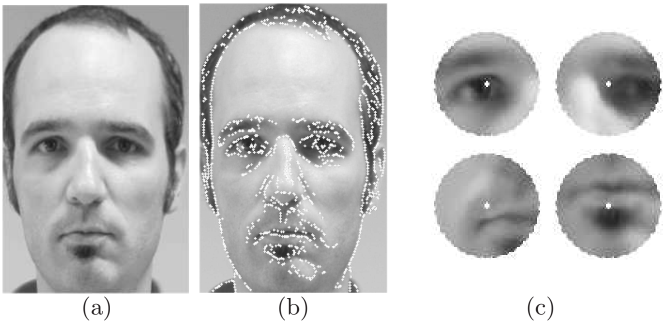


Fig. 5. Log polar sampling: (a) original image (b) all fixations (c) some reconstructed log-polar patches

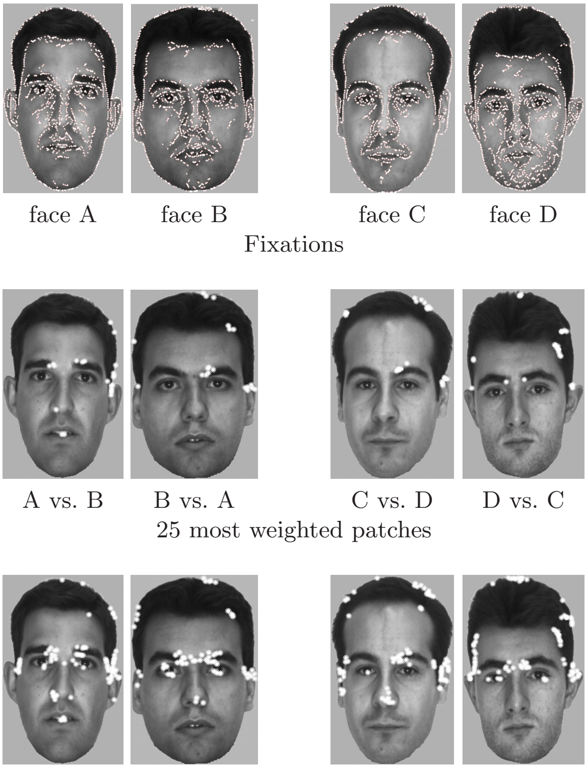
An interesting approach more related to this work extracts most salient patches (there denoted *fragments*) of a set of images [46]. There a sufficient coverage of patches are extracted from a set of "client" images, before each patch is weighted in terms of its mutual information with respect to a selected set of classes. However, the optimality criterion there used to select the most relevant patches differs from ours. We use a *deterministic* criterion computing the distance from the "impostor" set, while they adopt a *probabilistic* criterion based on empirical estimation of probability function. In order to obtain a reliable estimate, their approach thus requires a considerably large training set.

Multi-scale patches extraction. From each face-image, candidate patches are extracted. These patches must be spatially distributed in a way to cover most of the face area. This methodology is similar to the one adopted in patch-based image classification [47,48,49,50] and image characterization [51]. Since face recognition requires to process information at different spatial resolutions, there may be an advantage in extracting candidate patches at multiple scales. In agreement with the analysis presented in a previous section, a space-variant, multi-scale image sampling is adopted. This allows to avoid two notable pitfalls: (a) blind analysis - whereby information revealed at one scale is not usefully available at other scales, and (b) repeated image processing - which would add to the overall computational expense. Each face-image is sampled using patches derived from a log-polar mapping [27], considering the resulting sampled vectors as our features.

As an example, Figure 5(b) shows the sampling points (corresponding to fovea fixations) of one face.

In particular, the face-image is re-sampled at each point following a log-polar scheme so that the resulting set of patches represents a local space-variant remapping of the original image, centered at that point.

Finding differences between face-pairs. Without loss of generality, we start by considering the two-face case, i.e. when client set and impostor set contain



face A face B face C face D

Fixations

A vs. B B vs. A C vs. D D vs. C

25 most weighted patches

A vs. B B vs. A C vs. D D vs. C

100 most weighted patches

Fig. 6. Two examples of differences extracted from pairs of images of different persons: (A,B) and (C,D)

only one face each. Later we examine how this process can be expanded to the multi-face case.

The main idea is that the patches from one face-image will tend to form their own cluster in the feature space, while those of the other face-image ought to form a different cluster—e.g. see Fig. 8. The “distinctiveness” of each patch can be related to its locus in feature space with respect to other faces. Any patches of the first face, found near loci of a second face can be considered less distinctive since they may easily be confused with the patches of that second face, and thus may lead to algorithmic misclassification. Conversely, a patch lying on the limb of its own cluster, that is most distant from any other cluster, should turn out to be usefully representative, and may thus be profitably employed by a classifier.

We formalize the degree of distinctiveness of each face patch by weighting it according to its distance from the projection of the other data-cluster. Patches with the highest weights are then interpreted as encoding the most important differences between the two face-images.

Qualitative examples. All images used in the experiments were gray-level, with resolution 320 × 200 pixels, and cropped in order to reduce the influence of the background. Fixations, or centers of the patch sampling process (edge-points), were computed using zero-crossings of a LoG filter. After a preliminary evaluation, log-polar patch resolution was set to 15 eccentricity steps (N_r), at each of which there were 35 receptive fields (N_a), with a 70% overlap along the two directions (O_r and O_a). This represents a reasonable compromise between fovea resolution and peripheral context. Some examples of log-polar patches, rebuilt from the log-polar representations, are shown on Fig. 5(c).

Fig. 6 represents the comparison between different individuals.

The first two columns (subjects A and B) reveal that the main differences are in the ears and in the eyebrows: this is clearly evidenced in row 3 that shows that the first 25 patches are located on the ear in the right part of the face and on the eyebrows. This result is re-enforced when adding patches (last row): note how the left ear is now highlighted.

4 Video-Based Face Image Analysis

Conversely to previous hypotheses 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. 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 [52]. The proposed architecture (sketched in figure 7) 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.

4.1 Relevance of the Time Dimension

As shown by Vaina et al. [10], 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.

The time dimension is involved also when unexpected stimuli are presented [1,11]. 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.

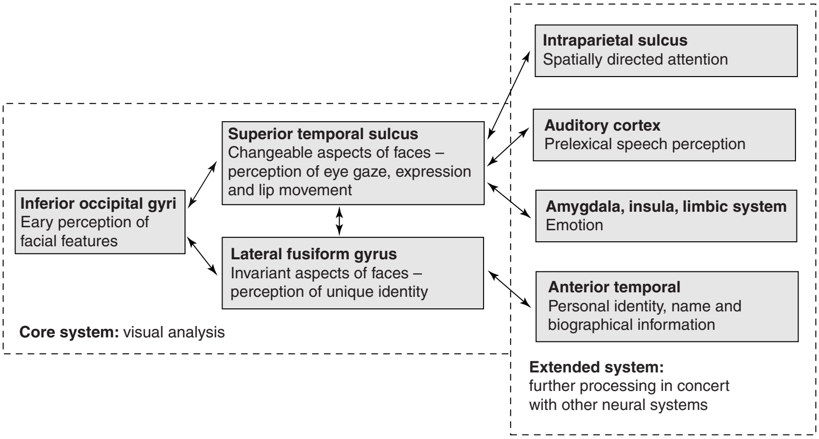

Intraparietal sulcus
Spatially directed attention
Auditory cortex
Prelexical speech perception
Superior temporal sulcus
Changeable aspects of faces – perception of eye gaze, expression and lip movement
Inferior occipital gyri
Eary perception of facial features
Amygdala, insula, limbic system
Emotion
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. 7. A model of the distributed neural system for face perception (reproduced from [52])

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.) [10,53,54,55,56,57,58]. As a consequence, non-rigid and idiosyncratic facial motions constitute a very powerful "dynamic template" which augments the information stored for familiar faces and may also improve the memory recall of structured information for identity determination [11].

4.2 A Computational Model for Computing Face Shape and Motion

The double layered architecture proposed by Haxby [52] 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 Markov Model* to emphasize the hierarchical nature of the process involved [59].

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 [60].

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

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 [61,62,63,64].

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.

   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

$$b(o|H_j) = \boldsymbol{\lambda}_j \tag{2}$$

   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 [66]). 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 [67]).

Verification of face sequences. Given few video sequences captured from the subject's face, the enrollment or modelling 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 $S = s_1, s_2, \cdots, s_T$, but 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 [61,64]. The resulting face HMM models are clustered in different groups based on their similarities [68,69]. 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 [66].
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 $S = s_1, s_2, \cdots, s_T$, using the Baum-Welch procedure and the emission probabilities found in the previous step (see Sect. 4.2). 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, modelling each of them with a spatial HMM. The expressions change during time is then modelled by the transition matrix of the PH-HMM, which constitutes the "temporal" model (as sketched in Fig. 8).

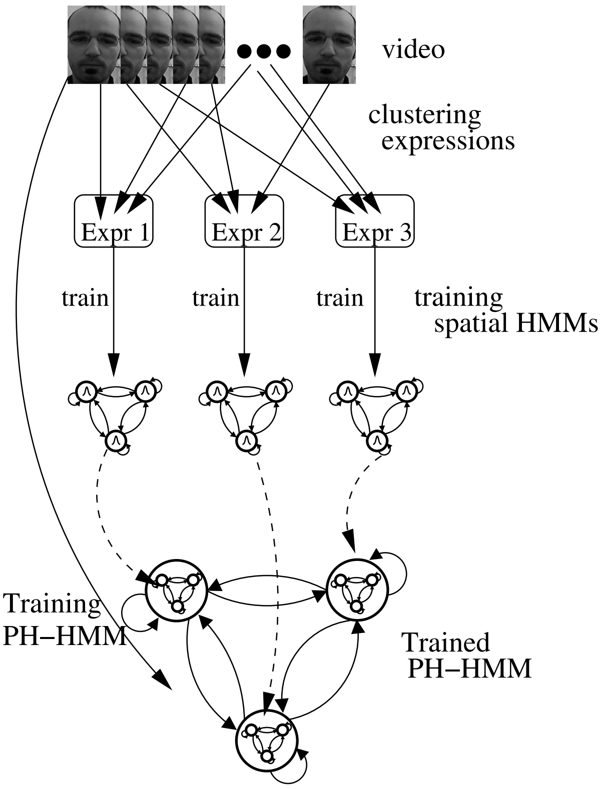


Fig. 8. Sketch of the enrollment phase of the proposed approach

4.3 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. Since each face is described with an HMM sequence, the expression clustering process is casted into the problem of clustering sequences represented by HMMs [68,69,70,71]. 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:

$$D(s_i, s_j) = \frac{P(s_j|\boldsymbol{\lambda}_i) + P(s_i|\boldsymbol{\lambda}_j)}{2} \tag{3}$$

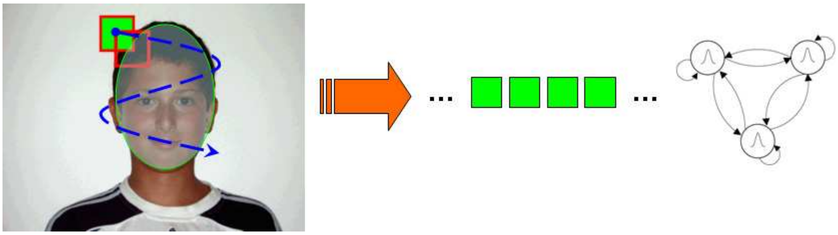


Fig. 9. Sampling scheme applied to generate the sequence of sub-images and the HMM model of the sampled sequence, representing a single face image

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 [68,69].

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 [72] has been used.

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) [66]. 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). It is defined as:

$$BIC(M_k) = \log P(X|\hat{M}_k) - \frac{1}{2}|\hat{M}_k| \log(N) \qquad (4)$$

where X is the data set (of cardinality N) to be modeled, $\{M_k\}$ ($k_{min} \leq k \leq k_{max}$) are the candidate models, $\hat{M}_k$ is the Maximum Likelihood estimate of the model M_k, and $|\hat{M}_k|$ is the number of free parameters of the model M_k.

4.4 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 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. The transition and the initial state matrices are estimated using the procedure described in section 4.2.

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.

4.5 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.

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 to 195 color images, with several changes in facial expression and scale (see fig. 10). The images have a resolution of 640x480 pixels. For the face classification experiments the images have been reduced to gray level with 8 bits per pixel. It is worth noting that there is no need for an explicit normalization for the different length of the sequences. The normalization in the time domain is obtained by self transitions of temporal HMM's states. In other words, if the subject takes 10 frames to change expression, it is likely that the system remains in the same expression state for 10 iterations before moving to the next state (self transitions).

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 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. Full client and impostor tests have been performed computing a ROC (Receiving Operating Characteristic) curve. 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.

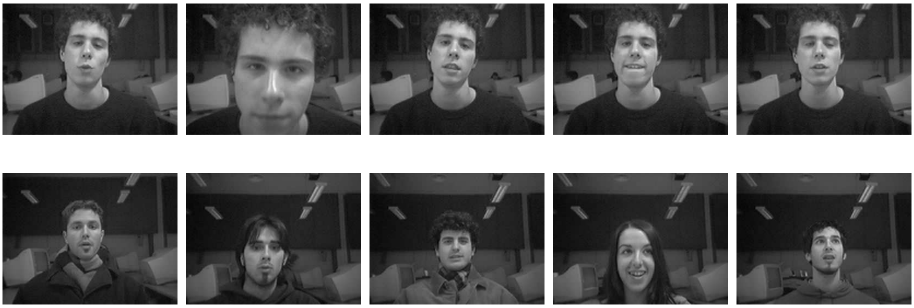


Fig. 10. (Top) Example frames of one subject extracted from the collected video database. (Bottom) One sample frame of five subjects, extracted from the first acquisition session.

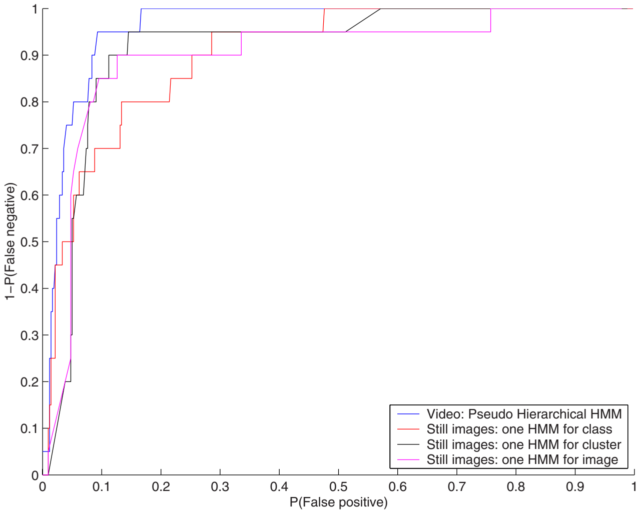


Fig. 11. The computed ROC curve for the verification experiment from video sequences of faces for the 4 methods reported

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. To choose the HMM that best fits the data, the Bayesian Inference Criterion (BIC) [66].

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.

The adopted 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

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%

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 [52] and described in section 4. 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.

5 Conclusions

The human visual system encompasses several complex mechanisms for parsing and analyzing the visual signal in space, time and frequency. These mechanisms, which include scale-space analysis and selective attention, allow the perception and recognition of complex and deformable objects, such as human faces. There is much to learn from the neural architecture of face perception and on the processes involved. Another important issue, which is rather difficult to address, is how human faces are "coded" in the brain. It seems that a complex mechanism exists which is adaptive to the nature of the perceived faces, i.e. if they are familiar or unfamiliar. Within this context, a crucial role is plaid by the concept of "model face", which is the reference for face detection and recognition. While a standard face model is required for distinguishing faces from non-faces, a personalized, user-dependent model is required for recognition. This concept can be stretched up to the definition of a subject-dependent face model, which is linked not only on the identification of standard facial landmarks, such as the eyes and the mouth (which indeed are demonstrated to be actively scanned by the gaze during face fixations) but rather on distinguishing face landmarks. These must correspond to very distinctive patterns on the face.

In this paper, a method to automatically extract the most distinguishing patterns in the subject's face has been proposed. The system, which has been tested on a standard face database, demonstrated to be able to select the face areas which are the most distinguishing for a given subject. The algorithm is based on the analysis of a number of randomly sampled matches on the face image. The results obtained show a remarkable similarity with the most prominent facial features perceived by human subjects. This method will be very important to devise facial templates which are not related to a general face model nor to a general template model, but rather the resulting template is fully adaptable to the subject's appearance.

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 [52]. 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.

Acknowledgments

This work has been supported by grants of the Italian Ministry of Scientific Research and by the IST European Network of Excellence Biosecure.

References

1. Knight, B., Johnston, A.: The role of movement in face recognition. Visual Cognition 4, 265–274 (1997)
2. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: Proc. Int. Conf. on Automatic Face and Gesture Recognition (1998)
3. Biuk, Z., Loncaric, S.: Face recognition from multi-pose image sequence. In: Proc. of Int. Symp. on Image and Signal Processing and Analysis (2001)
4. Li, Y.: Dynamic face models: construction and applications. PhD thesis, Queen Mary, University of London (2001)
5. Shakhnarovich, G., Fisher, J.W., Darrell, T.: Face recognition from long-term observations. In: Proc. of European Conf. on Computer Vision (2002)
6. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding 91, 214–245 (2003)
7. Liu, X., Chen, T.: Video-based face recognition using adaptive hidden markov models. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition (2003)
8. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition (2003)
9. Hadid, A., Pietikäinen, M.: 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)
10. Vaina, L.M., Solomon, J., Chowdhury, S., Sinha, P., Belliveau, J.W.: Functional Neuroanatomy of Biological Motion Perception in Humans. Proc. of the National Academy of Sciences of the United States of America 98(20), 11656–11661 (2001)
11. OToole, A.J., Roark, D.A., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Science 6, 261–266 (2002)
12. Darwin, C.: The expression of the emotions in man and animals. John Murray, London, UK (1965) (original work published 1872)
13. Goren, C., Sarty, M., Wu, P.: Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics 56, 544–549 (1975)
14. Walton, G.E., Bower, T.G.R.: Newborns form "prototypes" in less than 1 minute. Psychological Science 4, 203–205 (1993)
15. Fagan, J.: Infants' recognition memory for face. Journal of Experimental Child Psychology 14, 453–476 (1972)
16. de Haan, M., Nelson, C.A.: Recognition of the mother's face by 6-month-old infants: A neurobehavioral study. Child Development 68, 187–210 (1997)
17. Ballard, D.H.: Animate vision. Artificial Intelligence 48, 57–86 (1991)
18. Aloimonos, Y.: Purposize, qualitative, active vision. CVGIP: Image Understanding 56(special issue on qualitative, active vision), 3–129 (1992)
19. Tistarelli, M.: Active/space-variant object recognition. Image and Vision Computing 13(3), 215–226 (1995)
20. Schwartz, E.L., Greve, D.N., Bonmassar, G.: Space-variant active vision: definition, overview and examples. Neural Networks 8(7/8), 1297–1308 (1995)
21. Curcio, C.A., Sloan, K.R., Kalina, R.E., Hendrickson, A.E.: Human photoreceptor topography. Journal of Computational Neurology 292(4), 497–523 (1990)
22. Sandini, G., Metta, G.: Retina- like sensors: motivations, technology and applications. In: Secomb, T.W., Barth, F., Humphrey, P. (eds.) Sensors and Sensing in Biology and Engineering, Springer, Heidelberg (2002)

23. Burt, P.J.: Smart sensing in machine vision. In: Machine Vision: Algorithms, Architectures, and Systems, Academic Press, London (1988)
24. Tong, F., Li, Z.N.: The reciprocal-wedge transform for space-variant sensing. In: 4th IEEE Intl. Conference on Computer Vision, Berlin, pp. 330–334. IEEE Computer Society Press, Los Alamitos (1993)
25. Schwartz, E.L.: Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. Biological Cybernetics (25), 181–194 (1977)
26. Fisher, T.E., Juday, R.D.: A programmable video image remapper. In: Proceedings of SPIE, vol. 938, pp. 122–128 (1988)
27. Grosso, E., Tistarelli, M.: Log-polar Stereo for Anthropomorphic Robots. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1842, pp. 299–313. Springer, Heidelberg (2000)
28. Yarbus, A.L.: Eye Movements and Vision. Plenum Press, New York (1967)
29. Yeshurun, Y., Schwartz, E.L.: Shape description with a space-variant sensor: Algorithms for scan-path, fusion and convergence over multiple scans. IEEE Trans. on PAMI PAMI-11, 1217–1222 (1993)
30. Shepherd, J.: Social factors in face recognition. In: Davies, G., Ellis, H., Shepherd, J. (eds.) Perceiving and remembering face, pp. 55–79. Academic Press, London (1981)
31. Nahm, F.K.D., Perret, A., Amaral, D., Albright, T.D.: How do monkeys look at faces? Journal of Cognitive Neuroscience 9, 611–623 (1997)
32. Haith, M.M., Bergman, T., Moore, M.J.: Eye contact and face scanning in early infancy. Science 198, 853–854 (1979)
33. Klin, A.: Eye-tracking of social stimuli in adults with autism. In: NICHD Collaborative Program of Excellence in Autism, May 2001, Yale University, New Haven, CT (2001)
34. Tistarelli, M., Grosso, E.: Active vision-based face authentication. Image and Vision Computing: Special issue on Facial Image Analysis 18(4), 299–314 (2000)
35. Bicego, M., Grosso, E., Tistarelli, M.: On finding differences between faces. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 329–338. Springer, Heidelberg (2005)
36. Wiskott, L., Fellous, J.M., der Malsburg, C.V.: Face recognition by elastic bunch graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligence 19, 775–779 (1997)
37. Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N., Nuflo, F.: Modelling visual attention via selective tuning. Artificial Intelligence 78, 507–545 (1995)
38. Lindeberg, T.: Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. Int. Journal of Computer Vision 11(3), 283–318 (1993)
39. Koch, C., Ullman, S.: Shifts in selective visual-attention - towards the underlying neural circuitry. Human Neurobiology 4, 219–227 (1985)
40. Salah, A., Alpaydın, E., Akarun, L.: A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(3), 420–425 (2002)
41. González-Jiménez, D., Alba-Castro, J.: Biometrics discriminative face recognition through Gabor responses and sketch distortion. In: Marques, J.S., Pérez de la Blanca, N., Pina, P. (eds.) IbPRIA 2005. LNCS, vol. 3523, pp. 513–520. Springer, Heidelberg (2005)
42. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision 60(2), 91–110 (2004)

43. Penev, P., Atick, J.: Local feature analysis: a general statistical theory for object representation. Network: computation in Neural Systems 7(3), 477–500 (1996)
44. Li, S., Hou, X., Zhang, H.: Learning spatially localized, parts-based representation. Computer Vision and Image Understanding 1, 207–212 (2001)
45. Kim, J., Choi, J., Yi, J., Turk, M.: Effective representation using ica for face recognition robust to local distortion and partial occlusion. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(12), 1977–1981 (2005)
46. Ullman, S., Vidal-Naquet, M., Sali, E.: Visual features of intermediate complexity and their use in classification. Nature Neuroscience 5, 682–687 (2002)
47. Agarwal, S., Roth, D.: Learning a sparse representation for object detection. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 113–130. Springer, Heidelberg (2002)
48. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, vol. 2, p. 264 (2003)
49. Dorko, G., Schmid, C.: Selection of scale-invariant parts for object class recognition. In: Proc. Int. Conf. on Computer Vision, vol. 2, pp. 634–640 (2003)
50. Csurka, G., Dance, C., Bray, C., Fan, L., Willamowski, J.: Visual categorization with bags of keypoints. In: Proc. Workshop Pattern Recognition and Machine Learning in Computer Vision (2004)
51. Jojic, N., Frey, B., Kannan, A.: Epitomic analysis of appearance and shape. In: Proc. Int. Conf. on Computer Vision, vol. 2, pp. 34–41 (2003)
52. Haxby, J.V., Hoffman, E.A., Gobbini, M.I.: The distributed human neural system for face perception. Trends in Cognitive Sciences 4(6), 223–233 (2000)
53. Wiskott, L., Fellous, J.M., Kruger, N., von der Malsburg, C.: Face recognition and gender determination. In: Proceedings Int.l Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland, pp. 92–97 (1995)
54. Wechsler, H., Phillips, P., Bruce, V., Soulie, F., Huang, T. (eds.): Face Recognition. From Theory to Applications. NATO ASI Series F, vol. 163. Springer, Heidelberg
55. Cottrell, G., Metcalfe, J.: Face, gender and emotion recognition using holons. In: Touretzky, D. (ed.) Advances in Neural Information Processing Systems, San Mateo, CA, vol. 3, pp. 564–571. Morgan Kaufmann, San Francisco (1991)
56. Braathen, B., Bartlett, M.S., Littlewort, G., Movellan, J.R.: First Steps Towards Automatic Recognition of Spontaneous Facial Action Units. In: ACM Workshop on Perceptive User Interfaces, Orlando, FL, November 15-16, 2001, ACM Press, New York (2001)
57. Picard, R.W.: Toward computers that recognize and respond to user emotion. IBM System (39), 3/4 (2000)
58. Picard, R.W.: 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)
59. Bicego, M., Grosso, E., Tistarelli, M.: Person authentication from video of faces: a behavioral and physiological approach using Pseudo Hierarchical Hidden Markov Models. In: Zhang, D., Jain, A.K. (eds.) Advances in Biometrics. LNCS, vol. 3832, pp. 113–120. Springer, Heidelberg (2005)
60. Rabiner, L.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. of IEEE 77(2), 257–286 (1989)
61. Kohir, V.V., Desai, U.B.: Face recognition using DCT-HMM approach. In: AFI-ART. Proc. Workshop on Advances in Facial Image Analysis and Recogniti Technology, Freiburg, Germany (1998)

62. Samaria, F.: Face recognition using Hidden Markov Models. PhD thesis, Engineering Department, Cambridge University (October 1994)
63. Nefian, A.V., Hayes, M.H.: Hidden Markov models for face recognition. In: ICASSP. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Seattle, pp. 2721–2724 (1998)
64. Bicego, M., Castellani, U., Murino, V.: Using Hidden Markov Models and wavelets for face recognition. In: IEEE. Proc. of Int. Conf on Image Analysis and Processing, pp. 52–56. IEEE Computer Society Press, Los Alamitos (2003)
65. Bicego, M., Grosso, E., Tistarelli, M.: Probabilistic face authentication using hidden markov models. In: Proc. of SPIE Int. Workshop on Biometric Technology for Human Identification (2005)
66. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
67. Fine, S., Singer, Y., Tishby, N.: The hierarchical hidden markov model: Analysis and applications. Machine Learning 32, 41–62 (1998)
68. Smyth, P.: Clustering sequences with hidden Markov models. In: Mozer, M., Jordan, M., Petsche, T. (eds.) Advances in Neural Information Processing Systems, vol. 9, p. 648. MIT Press, Cambridge (1997)
69. Panuccio, A., Bicego, M., Murino, V.: A Hidden Markov model-based approach to sequential data clustering. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 734–742. Springer, Heidelberg (2002)
70. Rabiner, L., Lee, C., Juang, B., Wilpon, J.: HMM clustering for connected word recognition. In: ICASSP. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, pp. 405–408 (1989)
71. Li, C.: A Bayesian Approach to Temporal Data Clustering using Hidden Markov Model Methodology. PhD thesis, Vanderbilt University (2000)
72. Jain, A.K., Dubes, R.: Algorithms for clustering data. Prentice-Hall, Englewood Cliffs (1988)