PERGAMON

Pattern Recognition 33 (2000) 1161–1177

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

On internal representations in face recognition systems

Maxim A. Grudin*

Miros Inc., 572 Washington Street, Suite 18, Wellesley, MA 02482, USA

Received 28 September 1998; received in revised form 4 March 1999; accepted 4 March 1999

Abstract

This survey compares internal representations of the recent as well as more traditional face recognition techniques to classify them into several broad categories. The categories assessed include template matching and feature measurements, analysis of global and local facial features, and incorporation of interpersonal and intrapersonal variations of human faces. Analysis of the face recognition systems within those broad categories makes it possible to identify strong and weak sides of each group of methods. The paper argues that a fruitful direction for future research may lie in weighing information about facial features together with localized image features in order to provide a better mechanism for feature selection.

Keywords: Face recognition; Computer vision; Neural networks; Elastic graphs; Multiresolution techniques; Eigenfaces; Wavelets; Template matching; Principal component analysis

1. Introduction

Face recognition may seem an easy task for humans, and yet computerized face recognition systems still cannot achieve a completely reliable performance. The difficulties arise due to large variations in facial appearance, head size and orientation, and changes in environmental conditions. Such difficulties make face recognition one of the fundamental problems in pattern analysis.

Although computerized recognition of human faces was initiated more than 20 years ago, in the last decade there has been an explosion of scientific interest in this area. However, there is still no widely accepted benchmark for testing the developed systems. Therefore, comparison of different face recognition systems is no easy task. Comparison on the basis of their recognition performance is often misleading, since most of the systems are tested on different facial databases. Other factors that impact the performance are the accuracy of the face location stage and the number of actual face recognition techniques used in each system.

Over the last 10 years, there have been numerous reviews of face recognition techniques, with Samal and Iyengar [1], Valentin et al. [2], Chellappa et al. [3], and a recent issue of the IEEE Transactions on Pattern Analysis and Machine Intelligence [4], among the most prominent. This paper aims to update these previous surveys by reviewing many of the recent developments in this field. In addition, the present paper attempts to establish a set of underlying recognition principles used in the design of each reviewed technique. Thus, this paper differs from previous reviews by helping identify the strengths and weaknesses of each class of techniques as well as outlining a set of general principles that may find applications in future designs.

The remainder of this paper is organized as follows: Section 2 addresses the differences between the face recognition techniques that use feature measurements and those using template matching. Section 3 describes recognition of faces using comparison of whole faces rather than local features. It examines techniques of the principal component analysis, neural networks, and flexible templates. Face recognition methods that use localized features are presented in Section 4. Section 5 discusses methods that attempt to improve the recognition performance using intrapersonal variations of localized

* Corresponding author. Tel.: +781-235-0330x241; fax: +781-235-0720.

E-mail address: m.a.grudin@ieee.org (M.A. Grudin)

0031-3203/00/$20.00
PII: S0031-3203(99)00104-1

features. The discussion and conclusions are presented in Section 6.

2. A comparison between feature- and template-based models

The two traditional classes of techniques applied to the recognition of frontal views of faces are measurements of facial features and template matching. The first technique is based on extraction of relative positions and other parameters of distinctive features (Fig. 1). Typical geometrical features include (from Brunelli and Poggio [5]):

- eyebrow thickness and vertical position at the eye center position;
- a description of the eyebrows' arches;
- nose vertical position and width;
- mouth vertical position, width, height, upper and lower lips;
- radii describing the chin shape;
- face width at nose position;
- face width halfway between nose tip and eyes.

The measured features must be normalized in order to be independent of position, scale, and rotation of the face. A set of the above measurements is stored as a feature vector. Once obtained from the input image, the feature vector is compared with an existing database of the feature vectors to find the best match. Early attempts of feature-based face recognition included works of Bledsoe [6], Goldstein et al. [7], and Kaya and Kobayashi [8]. Some of the more recent investigations can be found in Refs. [9–12].

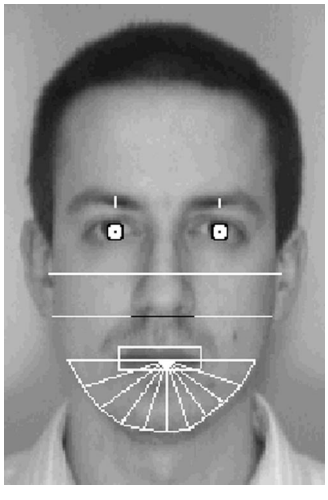


Fig. 1. Geometrical measurements used for feature-based recognition (from Brunelli and Poggio [5]).

In the simplest version of template matching, each person is represented as a database entry whose fields contain a two-dimensional array of pixels, extracted from a digital image of their frontal view. The image must be normalized in a similar fashion to that used in feature-based matching. Recognition is performed by comparison of the unclassified image with all the database images, using correlation as a typical matching function. Basic studies of the template-based matching were performed by Baron [13].

Comparison of these two classes of face recognition techniques was performed by Brunelli and Poggio [5]. The feature-based strategy showed a higher recognition speed and smaller memory requirements. However, it was concluded that the template-based technique is superior in recognition ratio. An increase in the number of measurements may improve recognition performance of the feature-based approach only slightly, because it is very difficult to improve the quality of the measurements. Also, the performance of the feature-based matching vastly deteriorates with partial face occlusions and any image degradations, such as camera misfocus. For the above reasons, our review will concentrate on the template-based techniques.

3. Extraction and analysis of global facial features

Out of the two most common directions in face recognition – analysis of global and local facial features – analysis of global features presumes a somewhat simpler problem formulation, since it avoids the question of selecting the size of localized features. Instead, images of the whole faces are aligned, typically in order to maximize the correlation between different facial images. In most cases, the alignment is performed with respect to the eye region, which is undisputedly the most discriminating area ([5], and many others). The faces are scaled so that the eyes in all faces correspond to the same physical locations.

3.1. Compact face representation using eigenvectors

One of the most well-known transformation applied to the facial images in order to extract global features is the *Principal Component Analysis* [14]. In this approach, a set of faces is represented by a small number of global eigenvectors, which encode the major variations in the input set. Originally, it was applied to faces by Sirovich and Kirby [15], who performed approximate reconstruction of faces in the ensemble using a weighted combination of eigenvectors (eigenpictures), obtained from that ensemble. The weights that characterize the expansion of the given image in terms of eigenpictures are seen as global facial features. In an extension of that work, Kirby and Sirovich [16] included the inherent symmetry of

faces in the eigenpictures (Fig. 2). The latter method produces slight improvements in the reconstruction of faces.

Turk and Pentland [17] used *eigenfaces* for face detection and identification. Fig. 3 shows an average face and the first 15 eigenfaces. Given the eigenfaces, each face is represented as a vector of weights. The weights are obtained by projecting the image into the eigenface components by a single inner product operation. The identification of the test image is done by locating the database entry, whose weights are closest (in Euclidean distance) to the weights of the image. Especially, large differences between the image weights and the database entries typically indicate absence of a face in the input image.

The authors reported 96% correct classification over lighting variations, 85% over orientation variations and 64% over size variations. The authors conclude that the robust performance of their system under different lighting conditions is caused by a significant correlation between images with differences in illumination (see also Ref. [3]). However, Zhang et al. [18] show that the performance of an eigenface-based technique deteriorates when lighting variations cannot be characterized as "very small".

Fig. 2. First nine eigenpictures, in order from left to right and top to bottom (from Kirby and Sirovich [16]).

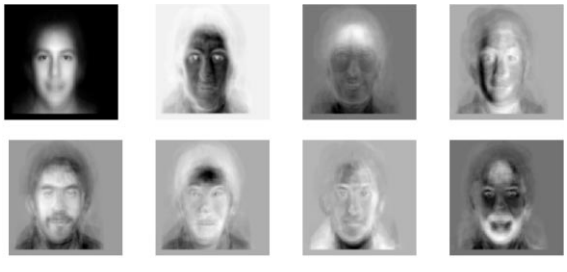


Fig. 3. The average face (top left corner) and eigenfaces (courtesy of A. Pentland).

Pentland et al. [19] extended the capabilities of their earlier system in several directions. Different structural configurations were considered, some of them included a search of facial features. The system showed 95% recognition rate on the FERET database, which contains 7562 images of approximately 3000 individuals. Using the system, the database can be interactively searched for images of certain types of people. To achieve orientation invariance, several entries with different head orientations are stored for each individual (see also Ref. [20]).

3.2. Properties of individual eigenvectors

O'Toole et al. [21]. studied the relationships between the values of facial eigenvectors and the characteristics of the faces, such as gender and race. It was shown that information in the weights of the second eigenvector yielded correct race predictions for 88.6% of the faces. Fig. 4 shows (from left to right) the first and the second eigenvectors, $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$, the sum of the first and the second eigenvectors, $\boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2$, resulting in a male image, and the result of subtracting the second eigenvector from the first one, $\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2$, resulting in a female image.

However, the most important information for face discrimination is found in the eigenvectors with smaller eigenvalues (Fig. 5). The eigenfaces with small eigenvalues contain information about higher frequencies, which also contains most of the image noise. Those eigenvectors are usually removed in favor of eigenvectors with the largest eigenvalues. The eigenvectors with large eigenvalues contain information about lower image frequencies, which contain less discriminative details of a face.

The authors [22,23] conclude that the strategy of minimizing the least-squares error is not the best one for the purposes of recognition. However, they do not provide a clear solution for overcoming the limitation of the approach based on eigenfaces, which we believe arises due to processing of whole faces rather than their constituent parts. In another study, Blackwell et al. [24] claimed that whole image preprocessing, such as PCA, *cannot* solve the problems associated with learning large, complicated data sets.

A more recent eigenface-based technique is described in Ref. [25]. The authors consider the class-conditional density as the most important object representation to be learned. Their maximum-likehood mechanism for face location uses two types of density estimates, a multivariate Gaussian for unimodal distributions and a Mixture-of-Gaussians model for multimodal distributions. Knowledge of those densities makes it possible to use a Bayesian framework for face recognition. The posterior

Fig. 4. Gender prediction using the second eigenvector (from O'Toole et al. [22]).

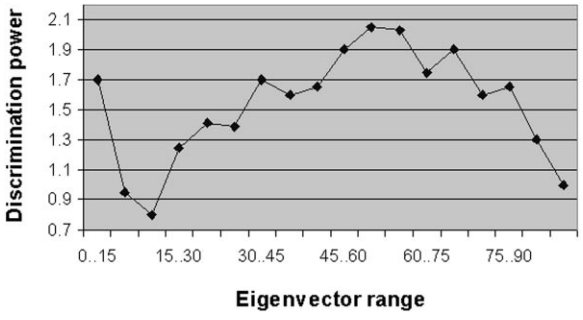


Fig. 5. Discrimination power as a function of the range of eigenvector (adopted from O'Toole et al. [22]).

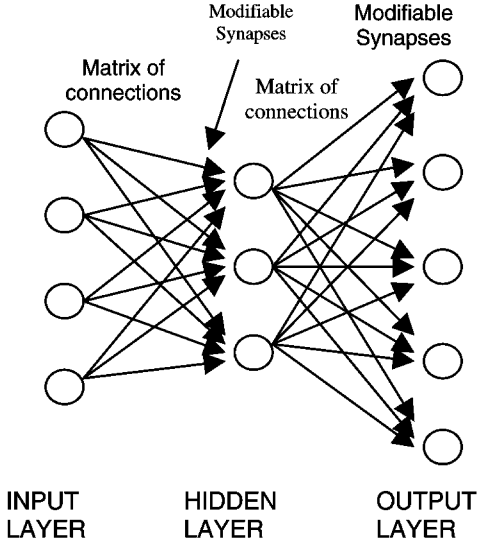


Fig. 6. A three-layer backpropagation network.

similarity measure is computed for the two classes corresponding to the intrapersonal and interpersonal variations. Some other approaches that take into account intrapersonal variations are described in later in this section and in Section 5.

3.3. Neural networks for face recognition

A number of research and commercial face recognition systems use neural networks. The variety of neural-network techniques used to recognize faces is enormous, which makes it impossible to describe them all in a single survey. In this section, we will consider face recognition using the multi-layer perceptrons (MLP), which was used by perhaps the largest number of researchers. Examples of applications of other neural architectures to face recognition are covered in Sections 4.2 and 5, and in other surveys, for example Ref. [2].

Originally formulated by Werbos [26], MLPs contain several fully interconnected layers of nonlinear neurons (Fig. 6). The connections between neurons contain weights, whose values determine the pattern space of the training patterns. The connection weights are adjusted by the backpropagation rule, which minimizes the error of the association.

The aim of face processing using the MLP is to develop a compact internal representation of faces, which is equivalent to feature extraction. Therefore, the number L of hidden neurons is less than in either input or output layers. That causes the network to encode inputs in a smaller-dimensional subspace that retains most of the important information. Oja [27] showed that the linear hidden neural units span the same space as the same number of principal components with the highest eigenvalues. The major difference between representations using the hidden neurons and those using the principal components is that in the former case the variance is evenly distributed across the hidden units [28].

The simplest form of using MLP for face recognition is to feed the facial image into the input layer without applying and preprocessing. Trying to retain only the most significant information, Kosugi [29] decreased

resolution of the facial image down to 12×12 pixels, which were fed into the MLP. Luebbers et al. [30] decomposed the image into a series of binary images using isodensity regions of the facial image [31]. Vincent et al. [32] trained several MLPs to locate facial features in the images (one network per feature). Five neural networks were used for each of the eye regions and two for the corners of the mouth.

Systems that directly associate the pixel information with a high-level syntactic description of an object are extremely sensitive to changes in the image. More sophisticated techniques extract features before feeding them into the network. Goudail et al. [33] used local autocorrelations for regions of 3×3 to 11×11 pixels. The system proposed by Augusteijn and Skufca [34] used second-order statistical information about textural regions to classify the facial features.

The hidden units of the MLP contain information which can be used to classify input images according to their typicality, sex and identity [35]. In another implementation, Golomb et al. [36] use a cascade of two MLP's for gender classification (Fig. 7). Both MLPs consist of three neuronal layers. The compression network is trained to reconstruct faces using a compact representation of the hidden layer. Once the weights in the compression network have reached equilibrium, the values of the hidden neurons are fed into the second MLP, which is trained to associate a person with its gender. The method produced 91% correct performance on 10 new faces.

In summary, the MLP approach has a similar representation to the approach based on eigenfaces. Lanitis et al. [37] point out that an important difference of these two approaches arises from the fact that the internal representation of the MLP is created during a training stage, which is specific for each particular application. In the application to face recognition, the researchers tend to associate different appearances of a single face with the person's identity. Considering faces as points in the decision space, the network learns to reduce the distance between different appearances of the same person while increasing the distance between faces of different people.

3.4. Recognition using flexible models

Another implementation that uses eigenvectors for face recognition was developed by Lanitis et al. [38]. This approach is based on the use of flexible models, related to those proposed by Yuille et al. [39]. Flexible models consider a facial image as a 3-D projection of a visual object that belongs to a certain class. These models are allowed to translate, rotate and deform to fit the best representation of their shape present in the image.

The approach of flexible appearance models [38] consists of two phases – modeling, in which flexible models of facial appearance are generated; and identification, in which these models are applied for classifying images (Fig. 8). Overall, three models are used, describing shape variations, localized intensity profiles, and shape-free gray-level intensities (Section 3.5). Distributions of those parameters are learned during the modeling stage. At the recognition stage, shape parameters and gray-level information are used to compare the face to all the database entries.

All the models used in the system have the same mathematical form. They are generated by performing the principal component analysis on the training samples. The authors use discriminant analysis [40] to isolate intrapersonal variation and interpersonal variation. Some of the significant modes of shape variation account only for intrapersonal variation. Fig. 9(a) illustrates the effect of the four most significant modes of the intrapersonal variation. Notice that the first three modes just

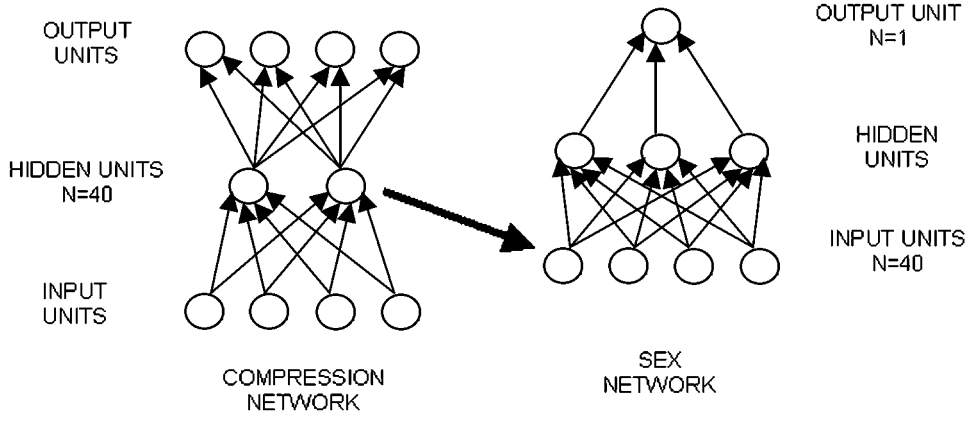


Fig. 7. Architecture of SexNET (adopted from Golomb et al. [36]).

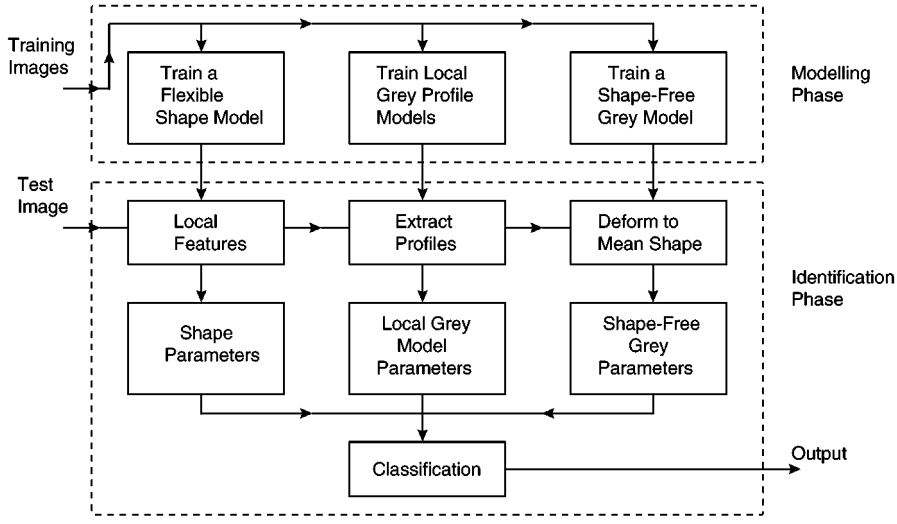


Fig. 8. Block diagram of the face identification system. Adopted from Lanitis et al. [38]).

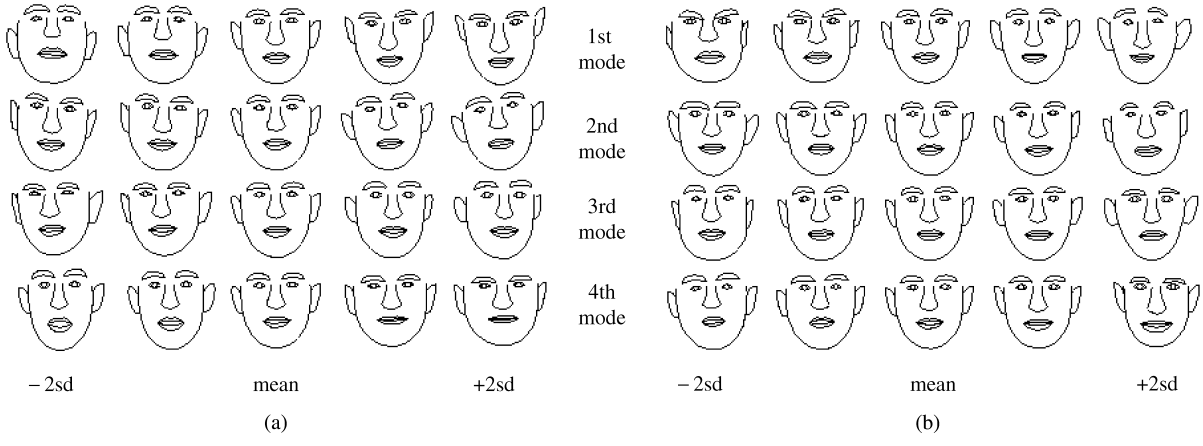


Fig. 9. The effect of: (a) main modes of the intrapersonal shape variation; (b) main discriminant modes of the interpersonal shape variation (from Lanitis et al. [38]).

change the 3-D orientation of the model, highlighting the importance of estimating the correct head orientation.

Fig. 9(b) shows the effect of varying the main four discriminant modes of interpersonal variation. Only six discriminant variables were needed to explain 95% of the interpersonal variation. Using those modes of variation, a new image is assigned to the class that minimizes the Mahalanobis distance D_{Ma} between the centroid of that class and the calculated appearance parameters.

In the practical realization, an active shape model [41] automatically locates a face in a new image. Once the model is fitted, both discriminant shape variables and gray-level parameters are measured. The obtained set of the appearance parameters is used to identify the person. A peak recognition performance of 95.5% was achieved on images from the test set of the Manchester Face Database, which contains images with variation in appearance, expression, 3-D head orientation, and scale.

3.5. Shape-free facial models

In their work on deformable facial templates, Lanitis et al. [38] distort facial images in order to achieve the best correspondence between the person's facial features and those of an abstract *shape-free* face (Fig. 10). The shape-free representation was originally proposed by

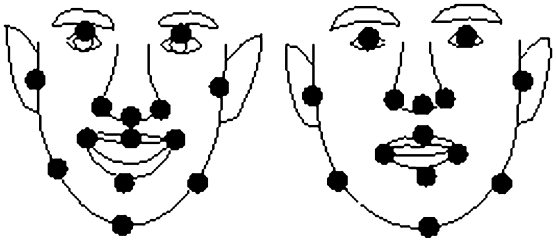


Fig. 10. Deformation of a facial image (left) to the shape-free face (right) using landmarks (from Lanitis et al. [38]).

Craw and Cameron [42]. They use the assumption that because the model of the principal component analysis is a linear space model, the faces themselves should form a linear space, that is the sum or average of two faces should itself be a face. Consequently, Craw [43] performs PCA of the faces that are preprocessed by distorting them to the shape-free form. More recently, the shape-free representation is used by other researchers (for example, Ref. [25]). The novelty of the approach by Lanitis et al. [38] is the utilization of the valuable shape information, which is used as another recognition cue (Section 3.4).

The shape-free form as proposed by Craw contains the distribution of gray-scale values. As an alternative, Grudin [44] proposes a shape-free model that contains a distribution of intrapersonal variations, which are related to the high-level facial features. Indeed, corresponding facial features of different people should exhibit similar intrapersonal variations. Humans use such information to identify a person, his/her emotions, and other characteristics. In the computerized recognition of human faces, estimated intrapersonal variations of facial features can be used to select the salient facial features from a single image of a person (Section 5.2.2).

4. Analysis of localized features

Many approaches that use whole face processing also integrate information about local features. In order to improve recognition performance, Moghaddam et al. [25] consider using eigenfeatures in addition to eigenfaces. The eigenfeatures of the eyes, the nose, and the mouth outperformed eigenfaces for a small number of principal components. In the system developed by Lanitis et al. [38], the local image profiles provided better recognition cues than the shape parameters.

We would like to distinguish between localized *image* features and localized *facial* features. Whereas facial features are composed of image features, they also use a priori knowledge about faces. This section describes two approaches that rely on analysis of localized image features. Initially, we present an approach that uses a technique similar to the PCA to analyze localized image features across a large number of facial images. As a result, the whole set of features is represented using a much smaller number of extracted features. The other described approach uses elastic links to preserve relational information between localized image features.

4.1. Local feature analysis

One of the biggest problems of utilizing localized features for the face recognition task is to select a subset of features that could reliably discriminate face in a large number of environments. In order to represent an image using a small number of local decorrelated features, Penev and Atick [45] proposed a technique called the *Local Feature Analysis* (LFA). The LFA produces a low-dimensional representation of visual objects that resembles the representation of the PCA. By enforcing the localization criterion, it becomes impossible to achieve perfect decorrelation between localized components; nonetheless, the reconstruction error for the LFA representation approaches that of the PCA representation.

As a result, local features are defined at each point in the image. However, there is still a significant residual correlation between such localized features. In order to reduce redundancy of the LFA representation, each localized population of the image features is represented by a single feature, while the rest of the features are suppressed. The resulting representation is sparse in a sense that the reconstruction of the most essential information in the image can be performed using a few features, which are distributed over the image. Fig. 11 shows the locations of the local features, the value of the topographic kernel and of a residual correlation for each of the localized features.

The LFA has been applied to face location and recognition. Fig. 12(a) illustrates utilization of the LFA to locate a face on a light uniform background. The local features used to locate the face are illustrated as dark dots. It is difficult to predict the method's performance in a cluttered scene, since this approach relies on the features located on the face outlines. In order to locate faces in cluttered background, some other studies use inner facial features (for example, Ref. [46]).

Fig. 12(b) illustrates 25 most prominent localized features that are used to recognize the facial image. Notice that many points are located near the head contour, and are therefore prone to noise due to in-depth head rotations or changes in expression. The change in the head orientation may also affect relative positions of the localized features that are located within the facial outlines. The authors propose to compensate this by recovering the 3-D head structure from a set of eigensurfaces. If such a 3-D structure can be recovered, it presents a significant advantage over recognizing faces in 2-D, since the shape

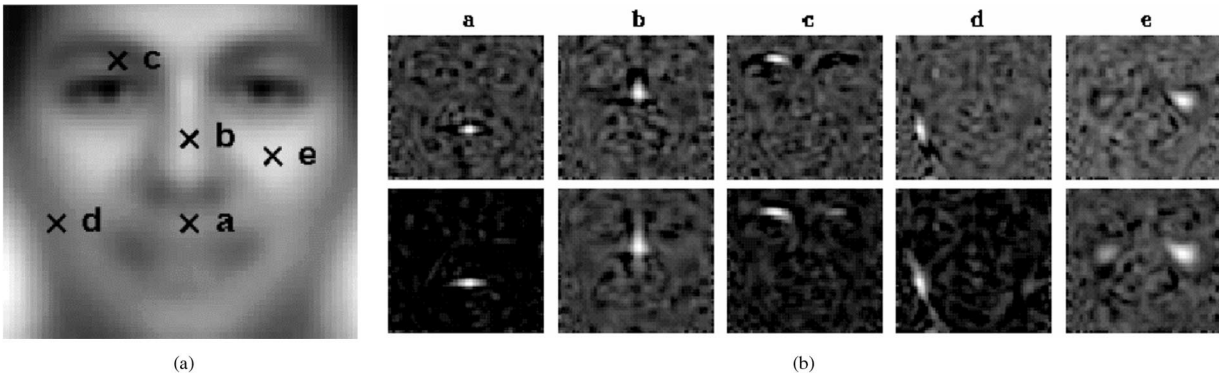


Fig. 11. (a) Locations of localized features; (b) representation kernel (top row) and residual correlations (bottom row) of the LFA (from Penev and Atick [45]).

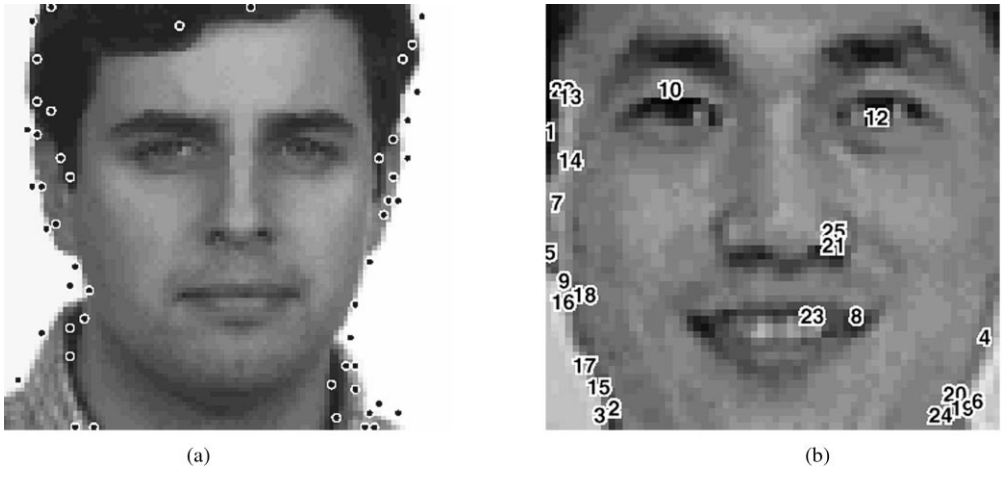


Fig. 12. (a) Face location using the LFA; (b) estimation of the most distinguishing features (from Penev and Atick [45]).

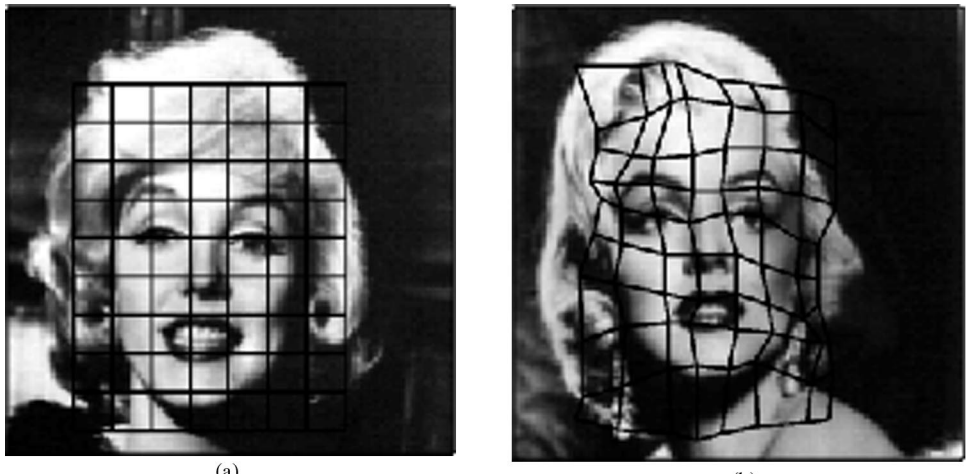
is independent of the image formation process. Another efficient approach to deal with changes in expressions and head rotations is described below.

4.2. Elastic graph matching

A problem that is central to recognition of faces as well as all other 3-D objects is preservation of the visual topology during the changes that arise due to different object projections or shape deformations. One of the effective solutions to this problem is the *Dynamic Link Architecture* (DLA) [47]. As designed, the DLA can be applied to recognition of most visual objects. But this approach has been applied to recognize faces and therefore it is described here in detail.

In a generic implementation, the DLA can be viewed as a regular grid, whose nodes contain a multiresolution description in terms of localized spatial frequencies. Each node contains several feature detectors that are based on modified Gabor-based wavelets [48]. Those detectors describe the gray-level distribution locally with high precision and more globally with lower precision. The grid nodes are connected with elastic links. Those connections group features into higher-order arrangements, which code for visual objects. The elasticity makes it possible to accommodate object distortions and changes in the viewing projection.

A new face is enrolled by manually positioning the grid over the face area (Fig. 13(a)). More than one graph may be stored for one person in order to accommodate different facial appearances. When a test image is presented, it is transformed into a grid of vectors, called an image domain. The image recognition is performed by matching all stored prototypic graphs to the image domain, the goal being minimization of the cost function between individual pairs of nodes. If the prototypic graph(s) of one person matches significantly better than all the other graphs, the face in the image is considered as recognized.



(a) (b)

Fig. 13. (a) Initial position and (b) deformation of the elastic graph (images courtesy of Rolf Würtz).

The graph matching is performed in two stages. Initially, the face is located in the image using the non-distorted grid. Once the grid is positioned over the face, its structure is deformed in such a way that each node achieves a minimum of the cost function within a certain vicinity around that node. The final cost of matching is a weighted combination of the cost of matching each node in the grid and of the amount of the grid deformation. The system thus generalizes over moderate changes in size and orientation (Fig. 13(b)).

The generality of the DLP has its downsides. Lades et al. [47] claimed that a face could be located using only the lowest frequency band of the prototypic graph. Yet the system is designed to compare all frequency bands each time a match is performed, even during the face location stage. In addition, bundling several different frequency bands into a single vector makes it more difficult to select salient features, since many features allow good differentiation only within a certain range of frequencies. To a large extent, these downsides of the DLA can be eliminated by changing the architecture so that it accommodates intrapersonal variations such as different appearances and lighting conditions [4]. Some of the systems that address those problems are presented in the next section.

5. Accommodation of the intra-class variations of the localized features

One of the difficulties in designing reliable face recognition systems is that different people often look more similar to each other when captured in the same conditions than the same face captured under very different conditions [4]. The problem of dealing with intrapersonal variations is addressed in many recent face recognition techniques. This section presents several techniques that use intrapersonal variations during the feature selection process. We start with a description of a mathematical model of a neural system, followed by a group of methods related to the DLA.

5.1. *Dynamically Stable Associative Learning* (*Dystal*)

One of the interesting systems that accommodates intrapersonal variations of local features is based on neurophysiological research. It is a computational model of the mechanisms identified in marine snail and rabbit hippocampus [49]. In that research, a network called Dynamically Stable Associative Learning (Dystal) investigates interactions between two inputs, namely the *Conditioned Stimulus* (CS) and the *UnConditioned Stimulus* (UCS).

A single layer Dystal network consists of a group of elements referred to as output units, equal in number to the number of components in the UCS vector. Each output unit receives input from a receptive field (a subset of CS inputs) and one (scaled) component of the UCS input vector (Fig. 14). The UCS can be a classification signal, or it can have the same size as the CS input to for the purpose of pattern completion.

All the patterns learned by Dystal are stored in a set of patches; however, each patch individually stores only a single association between a CS input and its associated UCS component. A patch is composed of: (1) a patch vector, the running average of CS input patterns that share similar UCS values; (2) the running average of similar UCS values; and (3) a weight that reflects the frequency of utilization of the patch. Unlike in other networks, the weight is not used in the computation of

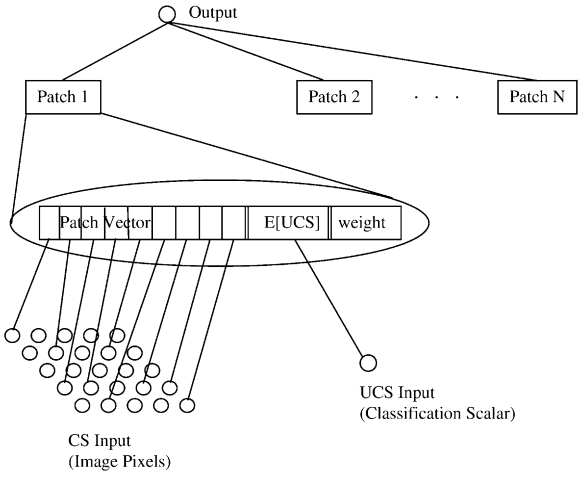


Fig. 14. Schematic representation of a single output neuron and its associated patches. Adopted from Alkon et al. [49].

the output of a neuronal element. The weight is used for patch merging and patch deletion. All patches whose weight has decayed to less than the patch retention threshold are removed.

Each output neural element computes the similarity function as the correlation factor between the CS pattern and each patch vector. The number and content of the patches are determined during training and are a function of the content of the training sets and of the global network parameters. During the training stage, the most similar patch is modified in order to increase its resemblance to the current CS pattern. As a result, each patch contains the average of all the CS inputs that are similar to that particular patch. As the training progresses, the effect that a novel CS input patch makes decreases.

Dystal has been applied to recognize hand-written postal index and hand-printed Japanese Kanji characters [24]. Prior to recognition, the digits were segmented, scaled and rotated to roughly the same orientation. The network was trained by presenting each pattern in the training set once. Dystal correctly classified 98% of previously unseen hand-written digits. When similarly trained to classify Kanji characters, it is able to learn 40 people's handprinting of 160 different characters to 99.8% accuracy. Such an approach might replace optical character recognition by optical word recognition, which is much faster due to reduced complexity of the segmentation problem.

When applied to face recognition, Dystal was able to correctly classify 100% of a small set of faces, which were exposed to variances in expression [50]. The faces need to be scaled and adjusted in their position prior to training and recognition. Dystal was trained on four presentations of each face and later tested on five other presentations. It was able to associate new appearances of the stored faces according to their most salient features. Significant changes in facial expressions did not affect the reconstruction of the original, e.g., changes in the mouth expression were ignored. This illustrates the ability of the network to concentrate on the stable features.

5.2. Graph-based techniques

Among methods that analyze intrapersonal variations perhaps the largest share belongs to methods inspired by the DLA (Ref. [51–55] and others). This occurred because the generic method of the DLA conveniently allows analysis the image data on multiple resolutions and elegantly accommodates invariance to object deformations. It is possible to outline two major directions of the development in this area. The first group of approaches is based on associating a small number of feature vectors with high-level facial features. The second group exploits the multiscale nature of image processing to reduce redundancy of the image representation.

5.2.1. The Topological Face Graph

One of the directions in the further research on attributed graphs is based on associating graph nodes with the high-level facial features [51,52]. That is, the same node corresponds to the same facial feature in different faces. As in the DLA, each node in such a graph contains image information on multiple resolutions. Different face orientations are encoded by graphs with different topology (Fig. 15).

All vectors in the *Face Bunch Graph* (FBG) referring to the same facial feature (called *fiducial point*) are bundled together in a bunch (Wiscott et al. [51]). Each fiducial point is represented by several alternatives in order to account for many possible variations in the appearance of that feature. Fig. 16 shows a sketch of an FBG. Each of the nine nodes is labeled with a bunch of six vectors, which together can potentially represent $6^{**}9 = 10077696$ different faces. When the FBG is matched to a face, a single vector (indicated in gray) that best encodes the appearance of the corresponding feature is selected from each bunch. The resulting image graph can be efficiently compared to large galleries without a need of repeated image search.

In the Krüger's approach, the feature vectors are also associated with facial features. However, only a single appearance is stored for each feature. The objective of that work is to evaluate typical discrimination abilities of different features. As a result, it is possible to design a similarity function that would assign different features certain weights, which are proportional to that feature's discrimination ability. Such a similarity function would

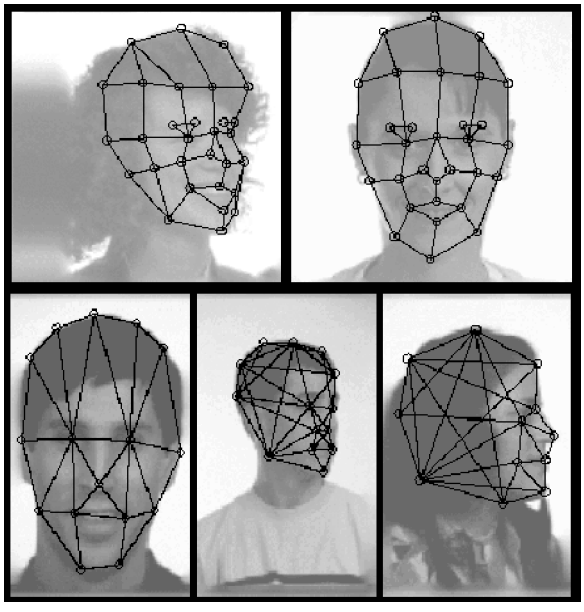


Fig. 15. Flexible graphs for different facial orientations and sizes (from Krüger [52]).

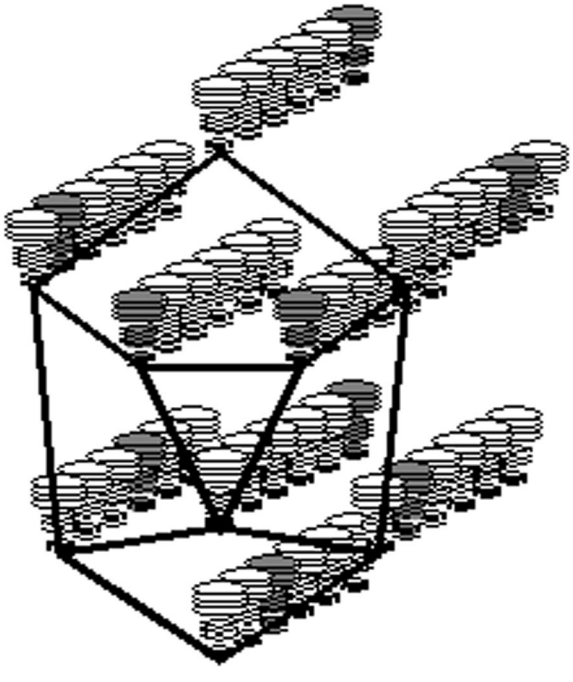


Fig. 16. Sketch of a face bunch graph (from Wiskott et al. [51]).

have a form:

$$Sim_{tot}(I,J) = \sum_{k=1}^{n} \beta_k \cdot Sim_{jet}(I_k, J_k), \quad (1)$$

where Sim_{jet} is the normalized dot product of the feature vectors (jets) in the nodes I_k and J_k, β_k is the significance weight of node k.

The approach was used for face recognition and pose estimation. The results, illustrated in Fig. 17, indicate that the weights for the discrimination problem (Fig. 17(a)) differ from the weights for the location problem (Fig. 17(b)). The eyes are more important for discrimination of frontal and half profile views than the mouth and chin. The nodes corresponding to the top of the head are very insignificant for pose estimation. The tip of the nose is the most significant feature, followed by the lips for the frontal and half-profile views and the chin for the profile view. The eyes were shown to be insignificant for that task.

5.2.2. Multiresolution analysis of facial images

Another major extension of the DLA is based on the idea that information on different image resolutions should be treated independently. Contributions in this direction were made by Würtz [53] and Grudin [54]. They use pyramidal architecture of the attributed graphs, which corresponds to the multiresolution nature of the image data. The sampling of each hierarchical grid is proportional to the size of the receptive fields in the nodes of that grid, in line with the requirements of the redundancy reduction [56].

The first of these approaches is motivated by (1) the necessity to reduce redundancy of the DLA structure; (2) the need to remove the background information, (3) use only the nodes that have correspondences in the input image. In that approach, each level of the graph is manually clipped so that the nodes whose receptive fields contain background information are removed. The nodes over the hair region are also removed in order to reduce dependency on the change in person's appearance.

In this multi-layered graph architecture, the nodes are linked using hierarchical and spatial links. The hierarchical links exist only between the parent and its children. The spatial links exist between the children of the same parent, forming a square if all four nodes are present. The matching is performed in a top-down manner. That is, the coarser level of the graph hierarchy is matched first, followed by matching the finer levels. Information from the coarse resolutions contributes to search on finer resolutions by setting the initial positions of the children relative to their parent. This reduces the computational complexity of matching the graph to the image and yet avoids the local minima of the cost function.

Würtz uses an assumption that only a subset of the graph nodes have a good correspondence in a new image of the same face. Some nodes do not have any correspondence at all, because the features they encode are changed or do not exist in the new image. Therefore, during the matching procedure, only the nodes with a

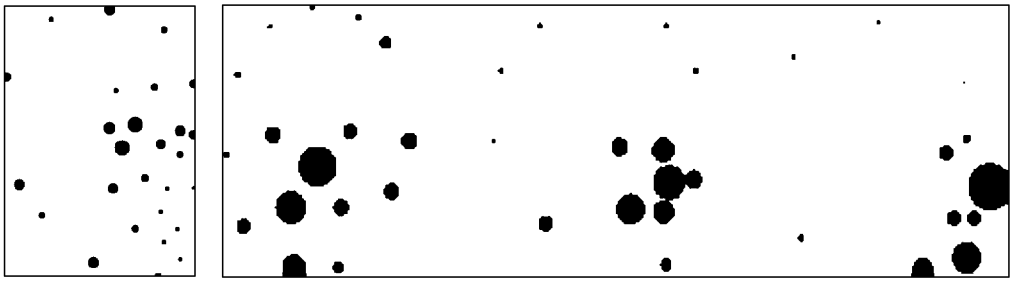


Fig. 17. (a) A weight matrix for the comparison of frontal with half profile views; (b) (left to right) the learned weight matrices for the pose estimation task for frontal (left), half profile (center), and profile (right) views (from Krüger [52]).

high degree of the correspondence are retained (Fig. 18). Such a feature selection approach can improve the system performance if the matching score of the correct attributed graph increases faster than those of the other graphs.

Although the Würtz's approach does not deal with intrapersonal variations in a direct manner, a closely related technique by Grudin [44] uses hierarchical graphs to estimate the distinguishing features from a single facial image. He uses the fact that humans can pick the most distinquishing facial features from a single image of a person. Humans use a priori knowledge to direct their attention to the areas that are very different between faces of different people and yet preserve certain predictable properties between different appearances of the same face.

This approach also uses a graph scheme, which differs from the previous approach in the implementation of the hierarchical structure and in the process of graph matching. Here, spatial links are used to connect direct neighbors at each level. This significantly improved the stability of the graph matching approach under considerable distortions [44]. In addition, the nodes on each level that have higher initial correspondence are matched before the others, thus providing anchors for the whole grid. Most of the previously implemented graph matching techniques relied on random or centrifugal sequence of matching the grid nodes.

Similarly to the previous approach, the nodes outside of the facial boundaries are manually removed during the enrollment stage. However, the nodes over the hair region are preserved, since they are expected to be automatically removed during the feature selection stage. The feature selection stage uses a Bayesian rule to estimate the discrimination confidence of individual localized features. The result of the feature selection is a sparse attributed graph. Each level in the graph contains features that are estimated to provide more discriminative information about the person's identity. The retained set of features is unique for each face.

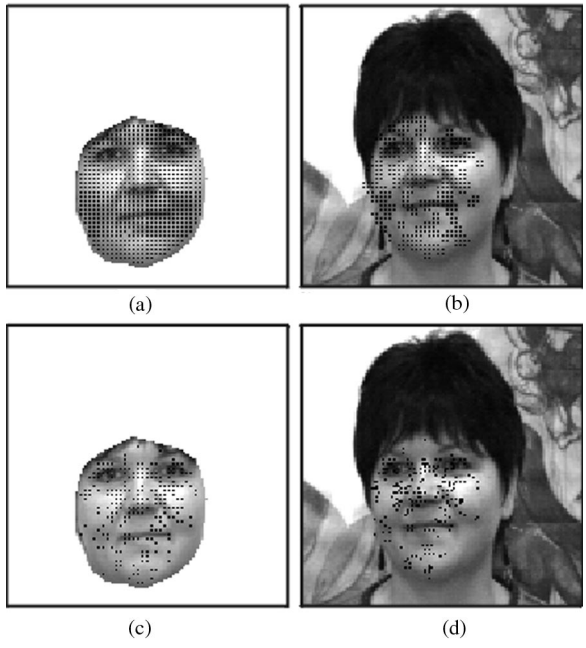


Fig. 18. (a) The original mapping of the high-resolution grid (left) and the distorted mapping over another appearance of the same face; (b) the retained good-matching nodes, shown over the original face and over the different face appearance (from Würtz [53]).

Ideally, a salient (or discriminative) feature should exhibit large interpersonal and low intrapersonal variations, with the value of the feature being very specific for the particular face. It is possible to estimate typical intrapersonal variations of the high-level facial features from a limited number of images. In practice, the distributions of the intrapersonal variations are computed for each person in the training set, mapped into the shape-free form, and averaged. The average shape-free distribution is subsequently used to estimate intrapersonal vari-

ations of new faces. To do that, the shape-free distribution is distorted to match shape characteristics of a particular face.

The experiment is performed in seven steps:

1. Obtain a single model graph from a person's image. Manually track a pre-defined set of facial features.
2. Compute the interpersonal distribution of image responses for each facial location.
3. For each person in the training set, match the person's model graph to other images of that person. Compute the mean and the standard deviation of the intrapersonal matching cost for each graph node.
4. Remap the distributions obtained in step 3 to a new representation, where positions of the facial features obtained in Step 1 correspond to pre-defined (shape-free) physical locations. Average the shape-free distributions obtained from the persons in the training set.
5. Remap the average shape-free distribution of intrapersonal variation so that the abstract facial features correspond to features of each particular person. The resulting new distribution contains intrapersonal variations that are related to the facial features of that person.
6. Use the Bayesian rule to compute the discrimination confidence for each node. Remove the nodes with lower confidence.
7. Match the sparsified model graphs to images in the test set.

In the hierarchical recognition scheme, the same facial region might exhibit different intrapersonal variation on different resolutions. Therefore, the distribution of intrapersonal variation on each processing scale is stored in a separate shape-free form. In Fig. 19, the lighter regions correspond to larger intrapersonal variations. The dark points in each image indicate positions of the eye centers and the mouth corners.

As illustrated in Fig. 19, some facial features exhibit significantly larger variations than others. On all scales, hair region exhibits large variations. The nose is also shown as less reliable, due to significant differences in its appearance under side-to-side head rotations. This occurs because the nose is the farthest feature from the spinal cord, which is the center of such rotations. At the same time, because the nose is the closest feature to the camera, its projections under different rotations are seen to contain more variations. The eyes and the mouth are seen as more stable features. However, their stability depends on the image resolution – for example, the iris movement on the high resolution contributes to higher intrapersonal variations of the eye region (Fig. 19(c)).

Fig. 20 illustrates the discrimination confidence of facial regions as computed by the Bayesian rule. The light regions correspond to more salient features. For illustrative purposes, the background areas are automatically filled in white. As a result of the feature selection stage, features with low discriminative power are removed, thus making the graph structure sparse.

When applied to the test set of the database, the sparse graphs recognized 85% of the facial images, compared with 78% for the non-sparse (complete) graphs. Fig. 21 illustrates examples of the graph matching on images taken in different conditions. Fig. 21(a) shows a complete graph being matched to the facial image. Fig. 21(b) shows the sparse graph matched to that image. In this and in the next image, the remaining nodes in the sparsified grid are illustrated as white squares. Fig. 21(c) illustrates the same sparse graph matched to another image of the

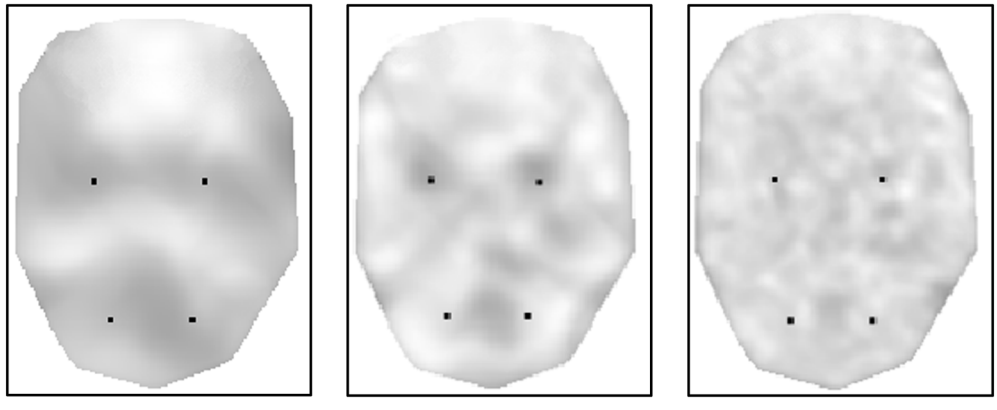


Fig. 19. The shape-free distributions of the intrapersonal variation at three resolutions (coarse to fine). Distribution (a) is obtained from low-resolution images, while the right image is obtained from images processed on a high resolution (From Grudin [44]). Large intrapersonal variations correspond to the lighter regions. The dark points indicate positions of the eyes and the mouth corners.

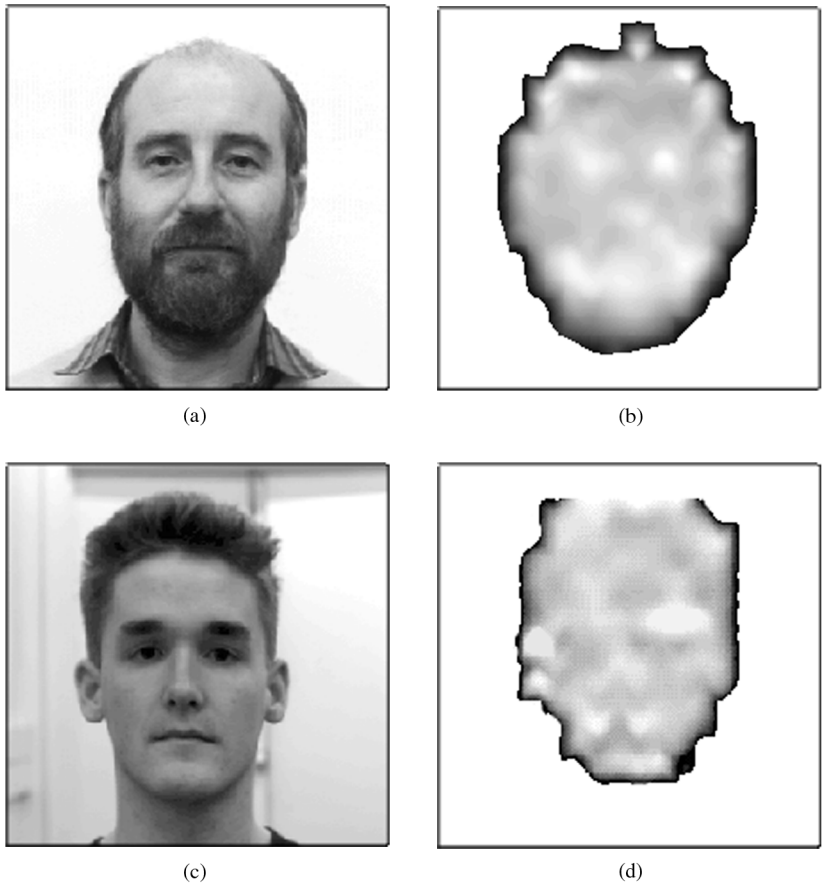


Fig. 20. Estimation of the discrimination power of feature descriptors (from Grudin [44]). The light regions correspond to more discriminative features.

same person. Fig. 21(d) shows the complete graph being matched to the difficult image of that person.

It is tempting to increase the sparsification ratio in order to improve the recognition performance while using fewer features. However, large values of the sparsification ratio increase the number of *loose* nodes, which do not have any adjacent neighbors and hence are not attached to the rest of the grid by the spatial links. If not removed, those loose nodes are not restricted in their movements over the image and introduce additional uncertainty in the matching score. Therefore, sparsification of the grid will improve the recognition performance only within a certain range of the sparsification factor.

Although the performance of this approach was worse than the performance of the approach by Lanitis et al. [37] when applied to the same test set of the Manchester face database, it was better than the performance of the best single recognition technique used in the system developed by Lanitis et al. [37]. However, although both techniques used the same face database, the performance cannot be compared directly due to different specifications of the test and training sets. Among the advantages of this technique is generation of a face model from single facial image. The major disadvantage is the necessity to upgrade the shape-free distribution of intrapersonal variation when a new set of environmental constraints, such as different lighting conditions, is introduced.

6. Discussion and conclusions

The complexity of the face recognition task makes it impossible for any single currently available approach to achieve 100% accuracy. Future successful face recognition systems will consist of multiple techniques, each being used to analyze a certain facial cue or a combination of cues. In such an implementation, the choice of the most appropriate technique will depend on the image context.

Although it is impossible to select a single best face recognition method, we can outline some guidelines that

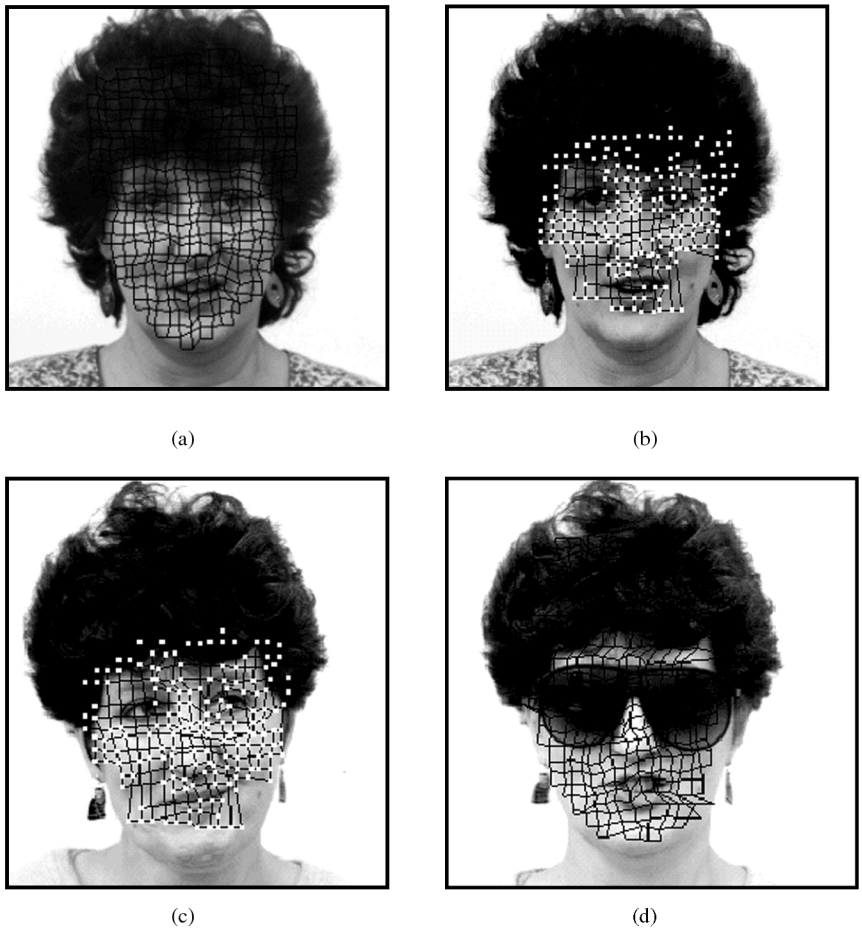


Fig. 21. Fitting of non-sparse (a, d) and sparse (b, c) grids to different images of the same person (from Grudin [44]).

are essential to achieve a high recognition ratio. The following discusses the set of principles that this paper has considered:

1. As was discussed in Section 2, the approaches based on feature measurements provide less reliability than the template-matching techniques. As a result, most of the recent research in face recognition was performed using template-based techniques.
2. In most cases, whole-face processing techniques achieve inferior performance compared to the methods that use localized features. One of the problems of processing the localized features is selecting the scale on which the features exhibit the most discriminating characteristics. Future face recognition systems will be addressing this problem on a regular basis.
3. The face recognition systems will be more sophisticated in terms of integrating the intrapersonal variations and reducing illumination dependency. At present, most of the systems are rather sensitive to such variations and these are the areas that might significantly improve the recognition accuracy.
4. Many existing techniques use localized *image* features, while some others use localized *facial* features. Although both might sometimes correspond to the same image region, there are certain differences between the two. If a system relies on a pre-defined set of localized facial features (e.g. nose or mouth), it neglects information about features that are specific to a particular person, such as birthmarks. On the other hand, if the systems that use localized image features do not incorporate information about the corresponding facial features, their performance would quickly degrade with changes in the image. Future designs will establish correspondence between the image features and the facial features in order to use a priori knowledge about facial features to select the most discriminating image features. In addition, future techniques are likely to consider how appearances of facial features

change under different transformations of facial images, such as in-depth head orientations and expressions.

5. Another cue that will be used more often in the future face recognition systems is shape information. Integration of such information will make it possible to compensate for the changes in facial expression. The shape characteristics may find greater usage in face recognition; in addition, they are likely to improve the image compression algorithms that are used in areas such as telecommunications.

Acknowledgements

The author wishes to acknowledge the research grant from the School of Engineering, Liverpool John Moores University. He would like to thank his supervisors Dr. David Harvey, Prof. Paulo Lisboa, and Dr. Mike (Showers) Shaw, and members of the Coherent and Electro-Optic Research Group (CEORG) in the Liverpool JMU for their constant support. He is grateful to Prof. Chris Taylor (University of Manchester), and Drs. Ben Dawson and James Kottas for their useful comments. The author would also like to thank Drs. Atick, Brunelli, Edwards, Krüger, O'Toole, Poggio, Pentland, Sirovich, Taylor, Wiskott, and Würtz for help with obtaining high-quality illustrations. Figs. 7, 8 and 10 are reprinted from *Image and Vision Computing*, Volume 13, A. Lanitis, C.J. Taylor, and T.F. Cootes, Automatic Face Identification System Using Flexible Appearance Models, 743–756, Copyright 1997, with permission of Elsevier Science.

References

[1] A. Samal, P.A. Iyengar, Automatic recognition and analysis of human faces and facial expressions: a survey, Pattern Recognition 25 (1992) 65–77.

[2] D. Valentin, H. Abdi, A.J. O'Toole, G.W. Cottrell, Connectionist models of face processing: a survey, Pattern Recognition 27 (1994) 1209–1230.

[3] R. Chellappa, C.L. Wilson, S. Sirohey, Human and machine recognition of faces: a survey, Proc. IEEE 83 (1995) 705–740.

[4] J. Daugman, Face and gesture recognition: overview, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 675–676.

[5] R. Brunelli, T. Poggio, Face recognition: features versus templates, IEEE Trans. Pattern. Anal. Mach. Intell 15 (1993) 1042–1052.

[6] W.W. Bledsoe. The model method in facial recognition, Panoramic Research Inc., Tech. Rep. PRI:15, Palo Alto, CA, 1964.

[7] A.J. Goldstein, L.D. Harmon, A.B. Lesk, Identification of human faces, Proc. IEEE 59 (1971) 748.

[8] Y. Kaya, K. Kobayashi, A basic study on human face recognition, in: S. Watanabe (Ed.), Frontiers of Pattern Recognition, Academic Press, New York, 1972, pp. 265–289.

[9] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (1990) 1481–1497.

[10] I. Craw, H. Ellis, J.R. Lishman, Automatic extraction of face features, Pattern Recognition Lett. 5 (1987) 183–187.

[11] M. Bichsel. Strategies of robust object recognition for identification of human faces. Ph.D. thesis, Eidgenossischen Technischen Hochschule, Zurich, 1991.

[12] X. Jia, M.S. Nixon, Extending the feature vector for automatic face recognition, IEEE Trans. Pattern. Anal. Mach. Intell 17 (1995) 1167–1176.

[13] R.J. Baron, Mechanisms of human facial recognition, Int. J. Man. Mach. Stud. 15 (1981) 137–178.

[14] K. Karhunen, Uber lineare methoden in der wahrscheinlichkeitsrechnung, Ann. Acad. Sci. Fennicae Ser. A1, Math. Phys. 37 (1946).

[15] L. Sirovich, M. Kirby, Low-dimensional procedure for the characterisation of human face, J. Opt. Soc. Amer. 4 (1987) 519–524.

[16] M. Kirby, L. Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Patt. Anal. Mach. Intell. 12 (1990) 103–108.

[17] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, Proceedings of the International Conference on Pattern Recoginition, 1991, pp. 586–591.

[18] J. Zhang, Y. Yan, M. Lades, Face Recognition: Eigenface, Elastic Matching and Neural Nets, Proceedings of the IEEE 85 (1997) 1423–1435.

[19] A.P. Pentland, B. Moghaddam, T. Starner, M.A. Turk, View-based and modular eigenspaces for face recognition, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1994, pp. 84–91.

[20] D.J. Beymer, Face Recognition Under Varying Pose, A.I. Memo No. 1461, MIT, New York, 1993.

[21] A.J. O'Toole, K.A. Deffenbacher, J. Barlett. Classifying faces by race and sex using an autoassociative memory trained for recognition, Proceedings. of the 13th Annual Conference Cognition Science Society, Hillsdale, Erlbaum, NJ; 1991.

[22] A.J. O'Toole, H. Abdi, K.A. Deffenbacher, D. Valentin, A low dimensional representation of faces in the higher dimensions of space, J. Opt. Soc. Amer. A 10 (1993) 405–411.

[23] A.J. O'Toole, H. Abdi, K.A. Deffenbacher, D. Valentin, A perceptual learning theory of the information in faces, in: T. Valentine (Ed.), Cognitive and Computational Aspects of Face Recognition, Routledge, London, 1995, pp. 159–182.

[24] K.T. Blackwell, T.P. Vogl, S.D. Hyman, G.S. Barbour, D.L. Alkon, A new approach to hand-written character recognition, Pattern Recognition 25 (1992) 655–666.

[25] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 696–710.

[26] P. Werbos. Beyond regression: new tools for prediction and analysis in the behavioral sciences. Ph.D. Thesis, Applied Mathematics, Harvard University, 1974.

[27] E. Oja, A simplified neuron model as a principal component Analyzer, J. Math. Biol. 13 (1982) 267–273.
[28] G.W. Cottrell, P. Munro. Principal component analysis of images via backpropagation. Proc. Soc. Photo-Optical Instrum. Engng. (1988) 1070–1076.
[29] M. Kosugi. Robust identification of human face using mosaic pattern and BNP, Proceedings of the International Conference on Neural Networks for Signal Processing, 1992, pp. 209–305.
[30] P.G. Luebbers, O.A. Uwechue, A.S. Pandya, A neural network based facial recognition system, Proc. SPIE 2243 (1994) 595–606.
[31] O. Nakamura, S. Mathur, T. Minami, Identification of human faces based on isodensity maps, Pattern Recognition 24 (1991) 263–272.
[32] J.M. Vincent, J.B. Waite, D.J. Myers, Automatic location of visual features by a system of multilayered perceptrons, IEE Proc.-F 139 (1992) 405–412.
[33] F. Goudail, E. Lange, T. Iwamoto, K. Kyuma, N. Otsu, Fast face recognition method using high order autocorrelations, Proceedings of the International Joint Conference on Neural Networks, 1993, pp. 1297–1300.
[34] M.F. Augusteijn, T.L. Skufca. Identification of human faces through texture-based feature recognition and neural network technology, Proceedings of the IEEE Conference on Neural Networks, 1993, pp. 392–398.
[35] G.W. Cottrell, M.K. Fleming, Face recognition using unsupervised feature extraction, Proceedings of the International Conference on Neural Networks. Paris, 1990, pp. 322–325.
[36] B.A. Golomb, D.T. Lawrence, T.J. Sejnowski, Sexnet: a neural network identifies sex from human faces, in: D.S. Touretzky, R. Lipmann (Eds.), Advances in Neural Computation Processing Systems, vol. 3, Kaufmann, San Mateo, 1991, pp. 572–577.
[37] A. Lanitis, C.J. Taylor, T.F. Cootes, Automatic interpretation and coding of face images using flexible templates, IEEE Trans. Pattern Anal. Machine Intell. 19 (1997) 743–756.
[38] A. Lanitis, C.J. Taylor, T.F. Cootes, Automatic face identification system using flexible appearance models, image and vision comput. 13 (1995) 393–401.
[39] A. Yuille, D. Cohen, P. Hallinan. Feature extraction from faces using deformable templates. Proc. IEEE Computer Society Conference on Computer Vision and Pattern. Recognition 1989 104–109.
[40] B.F.J. Manly, Multivariate Statistical Methods, a Primer, Chapman & Hall, London, 1986.
[41] T.F. Cootes, C.J. Taylor, A. Lanitis. Active shape models: evaluation of the multiresolution method for improving image search, Proceedings of the 5th Br. Machine Vision Conference. BMVA Press, 1994, pp. 327–336.
[42] I. Craw, P. Cameron. Parameterizing images for recognition and reconstruction, Proceedings of the BMVC 91 Glasgow, Scotland 1991, pp. 367–370.
[43] I. Craw, A manifold model of face and object recognition, in: T. Valentine (Ed.), Cognitive and Computational Aspects of Face Recognition: Explorations in Face Space, Routledge, London, 1995, pp. 183–203.
[44] M. Grudin, A compact multi-level model for the recognition of facial images Ph.D. Thesis Liverpool John Moores University, UK, 1997.
[45] P. Penev, J.J. Atick, Local feature analysis: a general statistical theory for object representation, Network: Comput. Neural Systems 7 (1996) 477–500.
[46] K.K. Sung, T. Poggio, Example-based learning for view-based human face detection, IEEE Trans. Pattern Anal. Machine Intell. 20 (1998).
[47] M. Lades, J.C. Vorbruggen, J. Buhmann, J. Lange, C.v.d. Malsburg, R.P. Wurtz, W. Konen, Distortion invariant object recognition in the dynamic link architecture, IEEE Trans. Comput. 42 (1993) 300–311.
[48] J. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by 2D visual cortical filters, J. Opt. Soc. Amer. (A) 2 (1985) 1160–1169.
[49] D.L. Alkon, Memory storage and neural systems, Sci. Am. 261 (1989) 42–50.
[50] D.L. Alkon, K.T. Blackwell, G.S. Barbour, S.A. Werness, T.P. Vogl, Biological plausibility of synaptic associative memory models, Neural Networks 7 (1994) 1005–1017.
[51] L. Wiskott, J.M. Fellows, N. Krüger, C.v.d. Malsburg, Face recognition by elastic bunch graph matching, IEEE Trans. on Pattern. Anal. Mach. Intell. 19 (1997) 775–779.
[52] N. Krüger, An algorithm for the learning of weights in discrimination functions using a priori constraints, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 764–768.
[53] R.P. Würtz, Object recognition robust under translations, deformations, and changes in background, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 769–774.
[54] M.A. Grudin, P.J.G. Lisboa, D.M. Harvey, Compact multi-level representation of human faces for recognition, Proceedings of IEE Conference on IPA-97 1997, 111–115.
[55] P. Kalocsai, H. Neven, J. Steffens. Statistical analysis of gabor-filter representation, Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 360–365.
[56] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells, J. Opt. Soc. Am. (A) 4 (1997) 2379–2394.

About the Author—DR. MAXIM GRUDIN received his Dipl. Eng. in Electrical Engineering from Vinnitsa State Technical University (Ukraine) in 1994. He received his Ph.D. degree from Liverpool John Moores University (United Kingdom) in 1997. His Ph.D. thesis is entitled "A Compact Multi-Level Model for the Recognition of Facial Images". At present, Dr. Grudin is a scientist at Miros, Inc., a Massachusetts-based company that develops security solutions based on face recognition technology. He has authored and co-authored ten papers and two patent applications.