# Face Recognition: A Literature Survey[1]

W. Zhao[2]
Sarnoff Corporation
R. Chellappa and A. Rosenfeld[3]
University of Maryland
P.J. Phillips[4]
National Institute of Standards and Technology

## Abstract

As one of the most successful applications of image analysis and understanding, face recognition has recently received significant attention, especially during the past several years. This is evidenced by the emergence of face recognition conferences such as AFGR [1] and AVBPA [2], and systematic empirical evaluations of face recognition techniques, including the FERET [3, 4, 5, 6] and XM2VTS [7] protocols. There are at least two reasons for this trend; the first is the wide range of commercial and law enforcement applications, and the second is the availability of feasible technologies after 30 years of research. This paper provides an up-to-date critical survey of still- and video-based face recognition research.

[1] The support of the Office of Naval Research under Grants N00014-95-1-0521 and N00014-00-1-0908 is gratefully acknowledged.

[2] Vision Technologies Lab, Sarnoff Corporation, Princeton, NJ 08543-5300.

[3] Center for Automation Research, University of Maryland, College Park, MD 20742-3275.

[4] National Institute of Standards and Technology, Gaithersburg, MD 20899.

## 1 Introduction

As one of the most successful applications of image analysis and understanding, face recognition has recently received significant attention, especially during the past few years. This is evidenced by the emergence of face recognition conferences such as AFGR [1] and AVBPA [2], and systematic empirical evaluations of face recognition techniques (FRT), including the FERET [3, 4, 5, 6] and XM2VTS [7] protocols. There are at least two reasons for this trend; the first is the wide range of commercial and law enforcement applications and the second is the availability of feasible technologies after 30 years of research.

The strong need for user-friendly systems that can secure our assets and protect our privacy without losing our identity in a sea of numbers is obvious. At present, one needs a PIN to get cash from an ATM, a password for a computer, a dozen others to access the internet, and so on. Although extremely reliable methods of biometric personal identification exist, e.g., fingerprint analysis and retinal or iris scans, these methods rely on the cooperation of the participants, whereas a personal identification system based on analysis of frontal or profile images of the face is often effective without the participant's cooperation or knowledge. The advantages/disadvantages of different biometrics are described in [8]. Table 1 lists some of the applications of face recognition.

| **Areas** | **Specific Applications** |
|---|---|
| Biometrics | Drivers' Licenses, Entitlement Programs |
| | Immigration, National ID, Passports, Voter Registration |
| | Welfare Fraud |
| Information Security | Desktop Logon (Windows NT, Windows 95) |
| | Application Security, Database Security, File Encryption |
| | Intranet Security, Internet Access, Medical Records |
| | Secure Trading Terminals |
| Law Enforcement and Surveillance | Advanced Video Surveillance, CCTV Control |
| | Portal Control, Post-Event Analysis |
| | Shoplifting and Suspect Tracking and Investigation |
| Smart Cards | Stored Value Security, User Authentication |
| Access Control | Facility Access, Vehicular Access |

Table 1: Typical applications of face recognition.

A general statement of the problem can be formulated as follows: Given still or video images of a scene, identify or verify one or more persons in the scene using a stored database of faces. Available collateral information such as race, age, gender, facial expression and speech may be used in narrowing the search (enhancing recognition). The solution to the problem involves segmentation of faces (face detection) from cluttered scenes, feature extraction from the face region, recognition or verification. In identification problems, the input to the system is an unknown face, and the system reports back the determined identity from a database of known individuals, whereas in verification problems, the system needs to confirm or reject the claimed identity of the input face.

Commercial and law enforcement applications of FRT range from static, controlled format photographs to uncontrolled video images, posing a wide range of different technical challenges and requiring an equally wide range of techniques from image processing, analysis, understanding and pattern recognition. One can broadly classify the challenges and techniques into two groups: static and dynamic/video matching. Within these groups, significant differences exist, depending on the specific application. The differences are in terms of image quality, amount of background clutter (posing challenges to segmentation algorithms), availability of a well-defined matching criterion, and the nature, type and amount of input from a user. In some applications, such as computerized aging, one is only concerned with defining a set of transformations so that the images created by the system are similar to what humans expect based on their recollections.

In 1995, a review paper by Chellappa et al. [9] gave a thorough survey of FRT at that time. (An earlier survey [10] appeared in 1992.) At that time, video-based face recognition was still in a nascent stage. During the past five years, face recognition has received increased attention and has advanced technically. Many commercial systems using face recognition are now available. Significant research efforts have been focused on video-based face modeling, processing and recognition. It is not an overstatement to say that face recognition has become one of the most successful applications of pattern recognition, image analysis and understanding.

In this paper we provide a critical review of the most recent developments in face recognition. This paper is organized as follows: In Section 2 we briefly review issues that are relevant from the psychophysical point of view. Section 3 provides a detailed review of recent developments in face recognition techniques using grayscale, range and other images. In Section 4 face recognition techniques based on video are reviewed, including face tracking, modeling, and non-face/face based recognition. Data collection and performance evaluation of face recognition algorithms are addressed in Section 5 with detailed descriptions of two representative protocols: FERET and XM2VTS. Finally, in Section 6 we discuss two difficult technical problems common to all the algorithms: lack of robustness to illumination and pose variations, and suggest possible ways to overcome these limitations.

## 2 Psychophysics/Neuroscience Issues Relevant to Face Recognition

In general, the human face recognition system utilizes a broad spectrum of stimuli, obtained from many, if not all, of the senses (visual, auditory, olfactory, tactile, etc.). These stimuli are used either individually or collectively for storage and retrieval of face images. In many cases contextual knowledge is also used, i.e. the surroundings play an important role in recognizing faces in relation to where they are supposed to be located. It is futile (using existing technology) to even attempt to develop a system that can mimic all these remarkable capabilities of humans. However, the human brain has its limitations in the total number of persons that it can accurately “remember”. A key potential advantage of a computer system is its capacity to handle large datasets of face images. In most applications the images are single or multiple views of 2-D intensity data, which forces the inputs to computer algorithms to be visual only. For this reason, the literature reviewed in this section is related to aspects of human visual perception.

Many studies and findings in psychology and neuroscience have direct relevance to engineers interested in designing algorithms or systems for machine recognition of faces. On the other hand, better machine systems can provide better tools for conducting studies in psychology and neuroscience [11]. For example, a possible engineering explanation of the lighting effect illustrated in [12] is as follows: for familiar faces a 3D model is usually built in memory; when the actual lighting direction is opposite to the usually assumed direction, a shape-from-shading algorithm recovers incorrect structural information and hence makes recognition of faces harder.

A complete review of relevant studies in psychophysics and neuroscience is beyond the scope of this paper. We only summarize findings that are potentially relevant to the design of face recognition systems. For details the reader is referred to the papers cited below. The issues that are of potential interest to designers are:

- **Is face recognition a dedicated process?** [13, 14]: Evidence for the existence of a dedicated face processing system comes from three sources [13]. A) Faces are more easily remembered by humans than other objects when presented in an upright orientation. B) Prosopagnosia patients are unable to recognize previously familiar faces, but usually have no other profound agnosia. They recognize people by their voices, hair color, dress, etc. Although they can perceive eyes, nose, mouth, hair, etc., they are unable to put these features together for the purpose of identification. It should be noted that prosopagnosia patients recognize whether the given object is a face or not, but then have difficulty in identifying the face. C) It is argued that infants come into the world prewired to be attracted by faces. Neonates seem to prefer to look at moving stimuli that have face-like patterns in preference to those containing no patterns or jumbled facial features. Some recent studies on this subject further confirm that face recognition is a dedicated process which is different from general object recognition [14]. Seven differences between face recognition and object recognition can be listed based on empirical results: 1) Configural effects (related to the choice of different types of machine recognition systems), 2) expertise, 3) differences verbalizable, 4) sensitivity to contrast polarity and illumination direction (related to the illumination problem in machine recognition systems), 5) metric variation, 6) rotation in depth (related to the pose variation problem in machine recognition systems), and 7) rotation in plane/inverted face.

- **Is face perception the result of wholistic or feature analysis?** [15] Both wholistic and feature information are crucial for the perception and recognition of faces. Studies suggest the possibility of global descriptions serving as a front end for finer, feature-based perception. If dominant features are present, wholistic descriptions may not be used. For example, in face recall studies, humans quickly focus on odd features such as big ears, a crooked nose, a staring eye, etc. One of the strongest pieces of evidence to support the view that face recognition involves more configural/holistic processing than other object recognition tasks has been the face inversion effect, where an inverted face is much harder to recognize than a normal face. An excellent example is given in [16] using the "Thatcher illusion" [17]. In this illusion, the eyes and mouth of a face are inverted. The result looks grotesque in an upright face; however, when shown inverted, the face looks fairly normal, and the inversion of the features is not readily noticed.

- **Ranking of significance of facial features:** Hair, face outline, eyes and mouth (not necessarily in that order) have been determined to be important for perceiving and remembering faces. Several studies have shown that the nose plays an insignificant role. In face recognition using profiles (which may be important in mugshot matching applications, where profiles can be extracted from side views), several fiducial points ("features") are in or near the nose region. Another outcome of some of the studies is that both external and internal features are important in the recognition of previously presented but otherwise unfamiliar faces, and internal features are more dominant in the recognition of familiar faces. It has also been found that the upper part of the face is more useful for face recognition than the lower part. The role of aesthetic attributes such as beauty, attractiveness and/or pleasantness has also been studied, with the conclusion that the more attractive the faces are, the better is their recognition rate; the least attractive faces come next, followed by the mid-range faces, in terms of ease of being recognized.

- **Caricatures** [18]: Perkins [19] formally defines a caricature as "a symbol that exaggerates measurements relative to any measure which varies from one person to another". Thus the length of a nose is a measure that varies from person to person, and may be useful as a symbol in caricaturing someone, but not the number of ears. Caricatures do not contain as much information as photographs, but they manage to capture the important characteristics of a face; experiments comparing the usefulness of caricatures and line drawings decidedly favor the former.

- **Distinctiveness:** Studies show that distinctive faces are better retained in memory and are recognized better and faster than typical faces. However, if a decision has to be made as to whether an object is a face or not, it takes longer to recognize an atypical face than a typical face. This may be explained by different mechanisms being used for detection and identification.

- **The role of spatial frequency analysis:** Earlier studies [20, 21] concluded that information in low spatial frequency bands plays a dominant role in face recognition. Later studies [22] showed that, depending on the recognition task, the low-, bandpass and high-frequency components may play different roles. For example the sex judgment task can be successfully accomplished using low-frequency components only, while the identification task requires the use of high-frequency components. The low-frequency components contribute to the global description, while the high-frequency components contribute to the finer details required in the identification task.

- **Viewpoint-invariant recognition?**[23, 24]: Much work in visual object recognition (e.g., [24]) has been cast within a theoretical framework introduced by Marr [25] in which different views of objects are analyzed in a way which allows access to (largely) viewpoint-invariant descriptions. Recently, there has been some debate about whether object recognition is viewpoint-invariant. In face recognition it seems clear that memory is highly viewpoint-dependent. Hill et al. [26] show that generalization even from one profile viewpoint to another is poor, though generalization from one 3/4 view to the other is very good.

- **Effect of lighting change**[12, 15, 27]: It has long been informally observed that photographic negatives of faces are difficult to recognize. However, relatively little work has explored why it is so difficult to recognize negative images of faces. In [12], experiments were conducted to explore whether difficulties with negative images of faces, and inverted images of faces, arise because each of these manipulations reverses the apparent direction of lighting, rendering a top-lit image of a face as if lit from below. This work demonstrated that bottom lighting does indeed make it harder to identity familiar faces. In [27], the importance of top lighting for face recognition, using the task of matching surface images of faces for identity, is demonstrated.

- **Movement and face recognition**[15, 28]: A recent intriguing study [28] shows that famous faces are easier to recognize when shown in moving sequences than in still photographs. This observation has been extended to show that movement helps in the recognition of familiar faces under a range of different types of degradations — negated, inverted, or thresholded (shown as black-and-white images) [15]. Even more interesting is that movement seems to provide a benefit even if the information content is equated in dynamic and static conditions. On the other hand, experiments with unfamiliar faces suggest no additional benefit from viewing animated rather than static sequences.

- **Facial expression**[29]: Based on neurophysiological studies, it seems that analysis of facial expressions is accomplished in parallel to face recognition. Some prosopagnosic patients, who have difficulties in identifying familiar faces, nevertheless seem to recognize facial expressions due to emotions. Patients who suffer from "organic brain syndrome" do poorly at expression analysis but perform face recognition quite well. Normal humans also exhibit parallel capabilities for facial expression analysis and face recognition. Similarly, separation of face recognition and "focused visual processing" tasks (look for someone with a thick mustache) has been claimed.

## 3 Face Recognition from Single Intensity or Other Images

In this section we survey the state of the art in face recognition in the engineering literature. Extraction of features such as the eyes and mouth, and face segmentation/detection are reviewed in Section 3.1. Sections 3.2 and 3.3 are detailed reviews of recent work in face recognition, including statistical and neural approaches.

## 3.1 Segmentation/detection and feature extraction

### 3.1.1 Segmentation/detection

Up to the middle 90's, most of the work in this area was focused on single-face segmentation from a simple or complex background. The approaches included using a face template, a deformable feature-based template, skin color, and a neural network. During the past five years, more reliable face detection methods have been developed to cope with multiple face detection in a complex background, where the face images may be partly occluded, rotated in plane, or rotated in depth. For technical details, please refer to [30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. Some of these methods were tested on relatively large databases, e.g. [30, 38]. A recent survey paper on face detection is [40]. Here we review two well-known approaches: The neural network approach of Kanade et al. [38, 39] and the example-based learning approach of Sung and Poggio [30]. A recent approach using a Support Vector Machine (SVM) is also briefly reviewed [37].

In [30], an example-based learning approach to locating vertical frontal views of human faces in complex scenes is presented. This technique models the distribution of human face patterns by means of a few view-based "face" and "non-face" prototype clusters. At each image location, a difference feature vector is computed between the local image pattern and the distribution-based model. This difference vector is then fed into a trained classifier to determine whether or not a human face is present at the current image location. The system detects faces of different sizes by exhaustively scanning an image for face-like local image patterns at all possible scales. More specifically, the system performs the following steps:

1. The input sub-images are all rescaled to size $19 \times 19$, and a mask is applied to eliminate near-boundary pixels. Normalization in intensity is done by first subtracting a best-fit brightness plane from the un-masked widow pixels and then applying histogram equalization.

2. A distribution-based model of canonical face- and non-face-patterns is constructed from samples. The model consists of 12 multi-dimensional Gaussian clusters; six of them represent face- and six represent non-face-pattern prototypes. The clusters are constructed by an elliptical $k$-means clustering algorithm which uses an adaptively varying normalized Mahalanobis distance metric.

3. A vector of matching measurements is computed for each pattern. This is a vector of distances between the test window pattern and the canonical face model's 12 cluster centroids. Two metrics are used; one is a Mahalanobis-like distance defined on the subspace spanned by the 75 largest eigenvectors of the prototype cluster, and the other is Euclidean distance.

4. A MLP classifier is trained for face/non-face discrimination using the 24-dimensional matching measurement vectors. The training set consists of 47316 measurement vectors, 4150 of which are examples of face patterns.

To detect faces in an image, preprocessing is done as in step 1, followed by matching measurement computation (step 3), and finally the MLP is used for detection. Results are reported on two large databases; the detection rate varied from 79.9% to 96.3% with a small number of false positives.

In [38], face knowledge is incorporated into a retinally connected neural network. The neural network uses image windows of size 20 × 20, and has one hidden layer with 26 units, where 4 units cover 10 × 10 non-overlapping subregions, 16 units cover 5 × 5 subregions, and 6 units cover 20 × 5 overlapping horizontal stripes. The image windows are preprocessed as described in step 1 above. To deal with overlapping detections, two heuristics are used: 1) "thresholding", where the classification of a face depends on the number of detections in a neighborhood, 2) "overlap elimination", where when a region is classified as a face, overlapping detections are rejected.

To further improve system performance, multiple neural networks are trained and their outputs are combined using an arbitrary strategy including ANDing, ORing, voting, or a separate arbitration neural network. A detection rate on a dataset of 130 test images varying from 77.9% to 90.3%, with an acceptable number of false positives, was reported. To handle faces at different angles, in [39] the authors propose using a router neural net to detect the angles of the faces. After angle detection, the virtual face detection system can be applied. The router neural network is a fully connected MLP with one hidden layer and 36 output units (each unit represents 10°).

In [37], a face detection scheme based on SVMs is proposed. SVM is a learning technique developed by Vapnik et al. at AT&T [41]. It can be viewed as a way to train polynomial, neural network, or Radial Basis Function classifiers. While most of the techniques used to train these classifiers are based on the idea of minimizing the training error, the *empirical risk*, SVMs operate on another induction principle, called *structural risk minimization*, which minimizes the upper bound of the generalization error. From an implementation point of view, training an SVM is equivalent to solving a linearly constrained Quadratic Programming (QP) problem. The challenge in applying SVMs to face detection is the complexity of solving a large scale QP problem. The authors propose using a decomposition algorithm to replace the original problem with a sequence of smaller problems. Their system is very similar to that in [30] except that no matching measurements are computed and the classifier is a SVM. The authors reported comparable results on two databases.

### 3.1.2 Feature Extraction

Feature extraction is the key to both face segmentation and recognition, as it is to any pattern classification task. For a comprehensive review of this subject see [9]. Here we review only a few representative techniques.

There has been renewed interest in the use of the Karhunen-Loeve (KL) expansion for the representation [42, 43] and recognition [44, 45] of faces. [42] considered the problem of KL representation of cropped face images. Noting that the number of images $M$ usually available for computing the covariance matrix of the data is much less than the row or column dimensionality of the covariance matrix, leading to singularity of the matrix, a standard method from linear algebra [46] is used that calculates only the $M$ eigenvectors that do not belong to the null space of the degenerate matrix. Once the eigenvectors (referred to as eigenpictures) are obtained, any image in the ensemble can be approximately reconstructed using a weighted combination of eigenpictures. By using an increasing number of eigenpictures, one gets an improved approximation to the given image. Examples of approximating an arbitrary image (not included in the calculation of the eigenvectors) by the eigenpictures are also given.

A generalized symmetry operator is used in [47] to find the eyes and mouth in a face. The motivation stems from the almost symmetric nature of the face about a vertical line through the nose. Subsequent symmetries lie within features such as the eyes, nose and mouth. The symmetry operator locates points in the image corresponding to high values of a symmetry measure discussed in detail in [47]. The procedure is claimed to be superior to other correlation-based schemes such as that of [48] in the sense that it is independent of scale or orientation. However, since no a priori knowledge of face location is used, the search for symmetry points is computationally intensive. A success rate of 95% is reported on a face image database, with the constraint that the faces occupy between 15-60% of the image.

A statistically motivated approach to detecting and recognizing the human eye in an intensity image with a frontal face is described in [49], which uses a template-based approach to detect the eyes in an image. The template has two regions of uniform intensity; the first is the iris region and the other is the white region of the eye. The approach constructs an "archetypal" eye and models various distributions as variations of it. For the "ideal" eye a uniform intensity is chosen for both the iris and whites. In an actual eye discrepancies from this ideal are present; these discrepancies can be modeled as "noise" components added to the ideal image. An $\alpha$-trimmed distribution is used for both the iris and the white, and the amount of degradation, which determines the value of $\alpha$, is estimated. $\alpha$ is easily optimized since the percentage of trimming and the area of the trimmed template are in 1-1 correspondence. A "blob" detection system is developed to locate the intensity valley caused by the iris enclosed by the white. In the experiments three sets of data were used. One consisted of 25 images used as a testing set, another had 107 positive eyes, and the third consisted of images with most probably erroneous locations which could be chosen as candidate templates. For locating the valleys, as many as 60 false alarms for the first data set, 30 for the second, and 110 for the third were reported A tabular representation of results for three sets of values for the $\alpha$'s is presented. An increase in the hit rate is reported when using an $\alpha$-trimmed distribution. The overall best hit rate reported was 80%.

[50] proposes an edge-based approach to accurately detecting two-dimensional shapes including faces. The motivations for proposing such a shape detection scheme are the following observations: 1) many two-dimensional shapes including faces can be well approximated by straight lines and rectangles, and 2) in practice it is more difficult to model the intensity values of an object and its background than to exploit the intensity differential along the object's boundary. Rather than looking for a shape from an edge map, edges are extracted directly from an image according to a given shape description. This approach is said to offer several advantages over previous methods of collecting edges into global shape description such as grouping and fitting. For example, it provides a tool for systematic analysis of edge-based shape detection. The computational complexity of this approach can be alleviated using multi-resolution processing.

To demonstrate the effectiveness of the proposed approach, results of face and facial

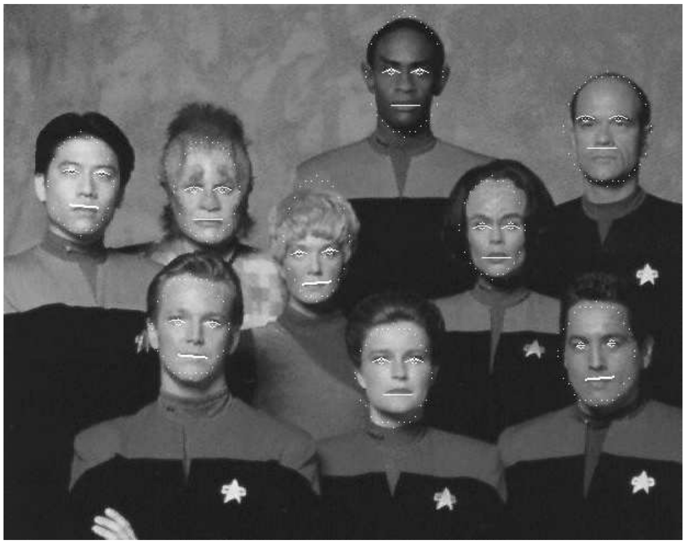


Figure 1: Face detection and facial feature detection in a group photo

feature detection are presented. One of these results is shown in Fig. 1 where the algorithm was applied to a group photo. For the detection of facial features, a small set of operators was designed. To limit the search space, the face center region is estimated using an ellipse-shaped operator, and is marked by a white dotted ellipse having the matched ellipse size. The face region detection is biased because only simple ellipses were fitted to the faces. Iris and eyelid detections are marked.

[51] presents a method of extracting pertinent feature points from a face image. It employs Gabor wavelet decomposition and local scale interaction to extract features at curvature maxima in the image. These feature points are then stored in a data base and subsequent target face images are matched using a graph matching technique. The 2-D Gabor function used and its Fourier transform are

$$
\begin{aligned}
g(x, y : u_0, v_0) &= \exp(-[x^2/2\sigma_x^2 + y^2/2\sigma_y^2] + 2\pi i[u_0 x + v_o y]) &(1)\\
G(u, v) &= \exp(-2\pi^2(\sigma_x^2(u - u_0)^2 + \sigma_y^2(v - v_0)^2)) &(2)
\end{aligned}
$$

where $\sigma_x$ and $\sigma_y$ represent the spatial widths of the Gaussian and $(u_0, v_0)$ is the frequency of the complex sinusoid.

The Gabor functions form a complete, though non–orthogonal, basis set. As with Fourier series, a function $g(x, y)$ can easily be expanded using the Gabor functions:

$$
\begin{aligned}
\Phi_\lambda(x, y, \theta) &= \exp\left[\,(-\lambda^2(x'^2 + y'^2)) + i\pi x'\right] &(3)\\
x' &= x\cos\theta + y\sin\theta &(4)\\
y' &= -x\sin\theta + y\cos\theta &(5)
\end{aligned}
$$

where $\theta$ is the preferred spatial orientation and $\lambda$ is the aspect ratio of the Gaussian.

The feature detection process uses a simple mechanism to model end-inhibition. It uses interscale interaction to group the responses of cells from different frequency channels. This results in the generation of the end-stop regions. The orientation parameter $\theta$

determines the direction of the edges. Hypercomplex cells are sensitive to oriented lines and step edges of short lengths, and their response decreases if the lengths are increased. They can be modeled by

$$I_{m,n}(x,y) = \max_{\theta} g\left( \parallel W_m(x,y,\theta) - \gamma W_n(x,y,\theta) \parallel \right) \tag{6}$$

and

$$W_j(x,y,\theta) = f \otimes \Phi(\alpha^j x, \alpha^j y, \theta), \qquad j = \{0, -1, -2, \cdots\} \tag{7}$$

where $f$ represents the input image, $g$ is a sigmoid non-linearity, $\gamma$ is a normalizing factor, and $n > m$. The final step is to localize the features at the local maxima of the feature responses.

Recently, the issue of feature detection accuracy has been addressed. In many systems, good recognition results are dependent on accurate feature (eyes, mouth) registration, and performance degradation is observed if the feature locations are not determined accurately enough [52]. [53] describes a robust and accurate feature localization method. In this method, images are pairwise registered using a robust form of correlation. The registration process is treated as an optimization problem in a search space defined by the set of all possible geometric and photometric transformations. At each point of the search space, a score function is evaluated and the optimum of this function is localized using a combined gradient-based and stochastic optimization technique. To meet real-time requirements and ensure high registration accuracy, a multiresolution scheme in used in both the image and parameter domains. After global registration, feature selection is based on minimizing the intra-class variance and at the same time maximizing the inter-class variance. Good results were obtained in experiments on a database (the extended M2VTS database [54]) containing 295 subjects.

## 3.2 Recognition from intensity images

### 3.2.1 Statistical Approaches

Eigenpictures (also known as eigenfaces) are used in [44] for face detection and identification. Given the eigenfaces, every face in the database can be represented as a vector of weights; the weights are obtained by projecting the image into eigenface components by a simple inner product operation. When a new test image whose identification is required is given, the new image is also represented by its vector of weights. The identification of the test image is done by locating the image in the database whose weights are the closest (in Euclidean distance) to the weights of the test image. By using the observation that the projection of a face image and a non-face image are quite different, a method of detecting the presence of a face in a given image is obtained. The method is illustrated using a large database of 2500 face images of sixteen subjects, digitized at all combinations of three head orientations, three head sizes and three lighting conditions. Several experiments were conducted to test the robustness of the approach to variations in lighting, size, head orientation, and the differences between training and test conditions. Impressive recognition rates were reported. It was also reported that the approach is fairly robust to changes in lighting conditions, but degrades quickly as the scale changes. One can explain this by the significant correlation between images obtained under different illumination conditions; the correlation between face images at different scales is rather low. Another way to interpret this is that the eigenfaces approach works well as long as the test image is "similar" to the ensemble of images used in the calculation of eigenfaces. The approach was also extended to real-time recognition of a moving face image in a video sequence. A spatio-temporal filtering step followed by a nonlinear operation is used to identify a moving person. The head portion is then identified using a simple set of rules and handed over to the face recognition module.

The capabilities of the system in [44] are extended in [45] in several directions. Extensive tests are reported based on 7562 images of approximately 3000 people. Twenty eigenvectors were computed using a randomly selected subset of 128 images. In addition to eigenrepresentation, annotated information on sex, race, approximate age and facial expression was included. Unlike mugshot applications, where only one front and one side view of a person's face is kept, in this database some of the persons are represented by many images with different expressions, headwear, etc.

More recently, practical face recognition systems have been developed based on eigenface representations. In [33], the eigenface method based on simple subspace-restricted norms is extended to use a probabilistic measure of similarity. The proposed similarity measure is based on a standard Bayesian analysis of image differences of two categories: 1) *intra-personal* variations in the appearance of the same individual due to different expressions or lighting, and 2) *extra-personal* variations in appearance due to difference in identity. The high-dimensional probability density functions for each class are then obtained from training data using an eigenspace density estimation technique and are subsequently used to compute a similarity measure based on the *a posteriori* probability of membership in the intra-personal class. Performance improvement of this probabilistic matching over the eigenface approach was demonstrated.

Face recognition systems using Linear/Fisher Discriminant Analysis [55] as the classifier have also been very successful [56, 57, 58, 59, 60, 61, 62, 63]. LDA training is carried out via scatter matrix analysis [64]. For an $M$-class problem, the within- and between-class scatter matrices $S_w$, $S_b$ are computed as follows:

$$S_w = \sum_{i=1}^{M} \Pr(\omega_i) C_i, \tag{8}$$

$$S_b = \sum_{i=1}^{M} \Pr(\omega_i)(\mathbf{m}_i - \mathbf{m}_0)(\mathbf{m}_i - \mathbf{m}_0)^T \tag{9}$$

where $\Pr(\omega_i)$ is the prior class probability and usually is replaced by $1/M$ in practice with the assumption of equal priors. Here $S_w$ is the within-class scatter matrix showing the average scatter $C_i$ of the sample vectors $\mathbf{x}$ of different classes $\omega_i$ around their respective means $\mathbf{m}_i$:

$$C_i = E[(\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T | \omega = \omega_i] \tag{10}$$

Similarly $S_b$ is the between-class scatter matrix, representing the scatter of the conditional mean vectors $\mathbf{m}_i$ around the overall mean vector $\mathbf{m}_0$. Various measures are available for quantifying the discriminatory power, a commonly used one being the ratio of the determinant of the between-class scatter matrix of the projected samples to the within-class scatter matrix of the projected samples:

$$\mathcal{J}(T) = \frac{|T^T S_b T|}{|T^T S_w T|}. \tag{11}$$

Let us denote the optimal projection matrix which maximizes $\mathcal{J}(T)$ by $W$; then $W$ can be obtained by solving the generalized eigenvalue problem [65]

$$S_b W = S_w W \Lambda_W \tag{12}$$

In [56], a face image retrieval system is reported based on discriminant analysis of the eigenfeatures, and in [57], a framework based on LDA for general object recognition is described. A general learning/recognition framework called SHOSLIF (Self-Organizing Hierarchical Optimal Subspace Learning and Inference Framework) is employed. SHOSLIF uses the theory of linear optimal projection to generate a hierarchical tessellation of a space defined by the training images. Using tree-structure learning, the eigenspace and LDA projections are recursively applied to smaller and smaller sets of samples. Such recursive partitioning is carried out for every node until the samples assigned to the node belong to a single class.

A comparative performance analysis was carried out in [58]. Four methods are compared: 1) a correlation-based method, 2) a variant of the linear subspace method suggested in [66], 3) an eigenface method [43, 44], and 4) a Fisher-face method which uses subspace projection prior to LDA projection to avoid the possible singularity in $S_w$ as in [56, 57]. Experiments were performed on a database of 500 images described in [67] and a database of 176 images created at Yale. The results show that the Fisher-face method performed significantly better than the other three methods. However, no claim is made about the relative performance of these algorithms on much larger databases.

To solve the generalization/overfitting problem when performing face recognition on a large face dataset but with very few training face images available per class, a holistic face recognition method based on subspace LDA was proposed [68](Fig. 2). Like existing methods [56, 58], this method consists of two steps: first the face image is projected into a face subspace via Principal Component Analysis (PCA), where the subspace dimension is carefully chosen, and then the PCA projection vectors are projected into the LDA to construct a linear classifier in the subspace. Unlike other methods, the dimension of the face subspace is fixed (for a given training set) regardless of the image size as long as the image size surpasses the subspace dimensionality. The property of relative invariance of the subspace dimension enables the system to work with smaller face images without sacrificing performance. This claim was supported by experiments using normalized face images of different sizes to obtain different face subspaces [62]. The choice of such a fixed subspace dimension is mainly based on the characteristics of the eigenvectors instead of the eigenvalues [60]. Such a choice of the subspace dimension enables the system to generate class-separable features via LDA from the full subspace representation, so that the generalization/overfitting problem can be addressed. In addition, a weighted distance metric guided by the LDA eigenvalues was employed to improve the performance of the subspace LDA method. The improved performance of generalized recognition was demonstrated on FERET datasets [63] and the MPEG-7 content set [69] in a proposal to MPEG-7 on robust face descriptors [70, 71]. A sensitivity test of the subspace LDA system is also reported in which an original face image is electronically modified by creating occlusions, applying Gaussian blur, randomizing the pixel location, and adding an artificial background. Figure 3 shows electronically modified face images which were correctly identified.

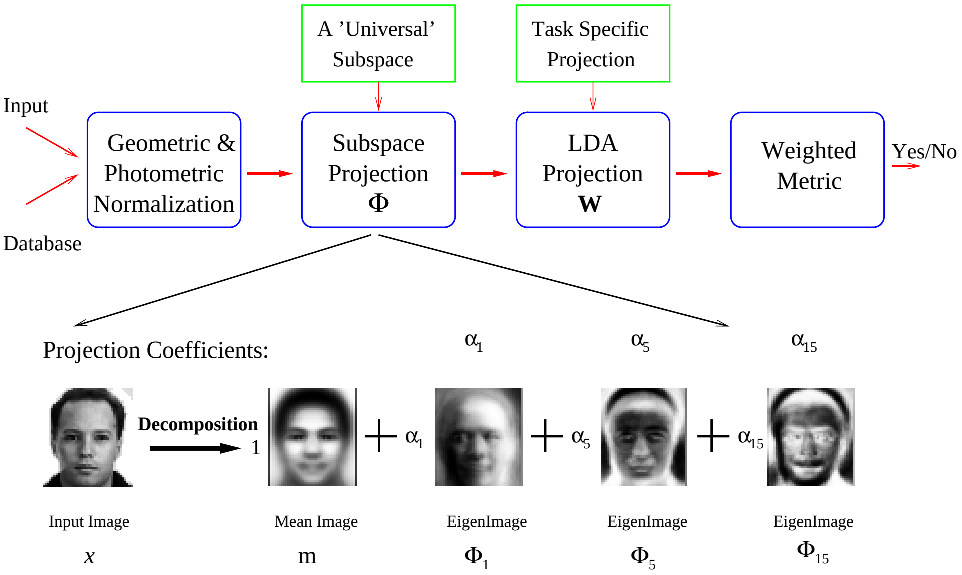


Figure 2: The subspace LDA face recognition system

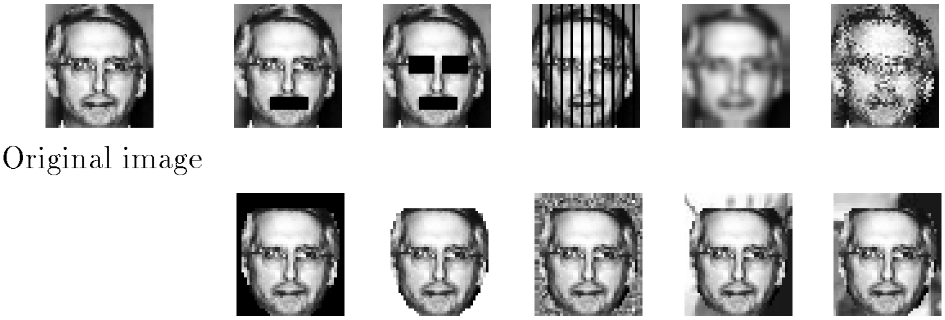


Figure 3: Electronically modified images that were correctly identified.

### 3.2.2 Neural Network Approaches

Neural networks (NN) have been used in face recognition to address several problems: gender classification, face recognition, and classification of facial expressions. One of the earliest demonstrations of NN for face recall applications used Kohonen's associative map [72]. Using a small set of face images, accurate recall was reported even when the input image was very noisy or when portions of the images were missing. This capability has also been demonstrated using optical hardware [73].

[74] describes an NN approach to gender classification using a vector of sixteen numerical attributes such as eyebrow thickness, widths of nose and mouth, six chin radii, etc. Two HyperBF networks [75] were trained, one for each gender. The input images were normalized with respect to scale and rotation by using the positions of the eyes, which were detected automatically. The 16-dimensional feature vector was also automatically extracted. The outputs of the two HyperBF networks were compared, the gender label for the test image being determined by the network with greater output. In the actual classification experiments only a subset of the 16-dimensional feature vector was used. The database consisted of 21 males and 21 females. The leave-one-out strategy [64] was employed for classification. When the feature vector from the training set was used as the test vector, 92.5% correct recognition accuracy was reported; for faces not in the training set, the accuracy dropped to 87.5%. Some validation of the automatic classification results has been reported using humans.

By using an expanded 35-dimensional feature vector, and one HyperBF per person, the gender classification approach has been extended to face recognition. The motivation for the underlying structure is the concept of a grandmother neuron: a single neuron (the Gaussian function in the HyperBF network) for each person. As there were relatively few training images per person, a synthetic data base was generated by perturbing the average of the feature vectors of available persons, and these persons were used as testing samples. For different sets of tuning parameters (coefficients, centers and metrics of the HyperBF's), classification results were obtained. Some corroboration of the caricatural behavior of the HyperBF networks, by psychophysical studies, was also presented.

The systems presented in [76] and [77] were based on the Dynamic Link Architecture (DLA). DLAs attempt to solve some of the conceptual problems of conventional artificial neural networks, the most prominent problem being the expression of syntactical relationships in neural networks. DLAs use synaptic plasticity and are able to instantly form sets of neurons grouped into structured graphs and maintain the advantages of neural systems. Both [76] and [77] used Gabor based wavelets for the features. A minimum of two levels, the image domain and the model domain, are needed for a DLA. The image domain corresponds to primary visual cortical areas and the model domain to the intertemporal cortex in biological vision. The DLA machinery is based on a data format able to encode information about attributes and links in the image domain and to transport that information to the model domain without including the image domain position. The structure of the signal is determined by three factors: the input image, random spontaneous excitation of the neurons, and interaction with the cells of the same or neighboring nodes in the image domain. Binding between neurons is encoded in the form of temporal correlations and is induced by the excitatory connections within the image. Four types of bindings are relevant to object recognition and representation: binding together all the nodes and cells that belong to the same object, expressing neighborhood relationships in the image of the object, bundling feature cells for features in different locations, and binding corresponding points in the image graph and model graph to each other. The DLA's basic mechanism, in addition to the connection parameter between pairs of neurons, is a dynamic variable ($J$) between pairs of neurons ($i$, $j$). $J$-variables play the role of synaptic weights for signal transmission. The connection parameters merely act to constrain the $J$-variables. The connection weights $J_{ij}$ are controlled by the signal correlations between neurons $i$ and $j$. Negative signal correlations lead to a decrease, and positive signal correlations to an increase, in $J_{ij}$. In the absence of correlation, $J_{ij}$ slowly returns to a resting state. Each stored image is presented by appropriately positioning a rectangular grid of points over the image and storing each grid point's locally determined jet. New image recognition takes place by mapping the image into the grid of jets and matching it to all the stored images. Conformation of the DLA is done by establishing and dynamically modifying links between the grid points.

The DLA architecture has been recently extended to Elastic Bunch Graph Matching [78, 79]. This is similar to the method described above, but instead of attaching only a single jet to each node, a set of jets is attached, each derived from a different facial image. To handle the pose variation problem in face recognition, the face pose is first determined using prior information [36] and the transformations of the sets under pose variation are learned [80]. Systems based on the EBGM approach have been applied in face detection, feature finding, pose estimation, gender classification, sketch image based recognition, and general object recognition. It is claimed that the success of the DLA/EBGM system may be due to its resemblance to the human visual system [14].

## 3.3 Other sensing modalities

### 3.3.1 Range Images

The discussion so far has considered only face recognition methods and systems that use data obtained from 2-D intensity images. Another topic being studied by researchers is face recognition from range image data. A range image contains the depth structure of the object in question. Although such data is not available in most applications it is important to determine the value of the added information present in range data in terms of its effect on the accuracy of face recognition.

[81] describes a template-based recognition system involving descriptors based on curvature calculations made on range image data. The data are obtained from a rotating laser scanner system with resolution better than 0.4mm. Surfaces are classified into planar, spherical, and surfaces of revolution. The data are stored in a cylindrical coordinate system as $f(\theta, y)$. At each point on the surface the magnitude and direction of the minimum and maximum normal curvatures are calculated. Since the calculations involve second-order derivatives, smoothing is required to remove the effects of noise in the image. This smoothing is done using a Gaussian filter.

Surface regions are classified as convex, concave and saddle. Ridges and valley lines are determined by obtaining the maxima and minima of the curvatures. The strategy used for face recognition is as follows:

- The nose is located.
- Locating the nose facilitates the search for the eyes and mouth.
- Other features such as forehead, neck, cheeks, etc. are determined by their surface smoothness (unlike hair and eye regions).

- This information is then used for depth template comparison. Using the locations of the eyes, nose and mouth the faces are normalized into a standard position. This position is re-interpolated to a regular cylindrical grid and the volume of space between the two normalized surfaces is used as the mismatch measure.

This system was tested on a dataset of 24 images of eight persons with three views of each. The data represented four male and four female faces. Adequate feature detection was achieved for 100% of these faces. 97% recognition accuracy was reported for the individual features and 100% for the whole face. In related work [82], the process of finding the features was formalized for recognition purposes.

### 3.3.2 Sketches and Infra-Red Images

In [83, 84], face recognition based on sketches, which are quite common in law enforcement, is described. Humans have a remarkable ability to recognize faces from sketches. This ability provides a basis for forensic investigations: an artist draws a sketch based on the witness's verbal description; then a witness looks through a large database of real images to determine possible matches. Usually, the database of real images is quite large, possibly containing thousands of real photos. Therefore, building a system capable of automatically recognizing faces from sketches has practical value. The first step in [83] is feature detection using deformable templates, applied to both the sketch image and the real photograph images. Then comes the key step of the system, photometric standardization. In this step, the pixels in the sketch image that have high intensity variations around facial features are replaced by Gaussian blurred versions, yielding so-called pseudo-images. Next, the pseudo-images and the real database images are geometrically standardized using a mesh face model. Finally, the eigenface method is used for classification. Recognition results are reported using 7 sketches and 16 photographs.

In [84], a system called PHANTOMAS (phantom automatic search) is described. This system is based on [77], where faces are stored as flexible graphs with characteristic visual features (Gabor features) attached to the nodes of the graph. The system was tested using a photo database of 103 persons (33 females and 70 males) and 13 sketches drawn by a professional forensic artist from the photo database. The results were compared with the judgments of five human subjects and were found to be comparable.

[85] describes an initial study comparing the effectiveness of visible and infra-red (IR) imagery for detecting and recognizing faces. One of the motivations in this paper is that changes in illumination can cause significant performance degradation for visible image based face recognition. Hence infra-red imagery, which is insensitive to illumination variation, can serve as an alternative source of information for detection and recognition. However, the inferior resolution of IR images is a drawback. Further, though IR imagery is insensitive to changes in illumination, it is sensitive to changes in temperature. Three face recognition algorithms were applied to both visible and IR images. The recognition results on 101 subjects suggested that visible and IR imagery perform similarly across algorithms, and that the fusion of IR and visible imagery is a viable means of enhancing performance beyond that of either alone.

So far we have not distinguished between two concepts: face identification and face verification. Strictly speaking, recognition includes both identification and verification.

In identification tasks, the input to the system is an unknown face, and the system reports its identity using a database of known individuals; whereas in verification tasks, the system needs to confirm or reject the claimed identity of the input face. To illustrate the difference between these two tasks, we will give a real example, the FERET evaluation, in Section 5.

### 3.4 Summary

Significant progress has been achieved in segmentation, feature extraction and recognition of faces in intensity images. As long as range images or stereo pairs are not available in commercial/law enforcement applications, face recognition can be viewed as a 2-D image matching and recognition problem with provisions for at most two or three views of each person's face.

In a recent comprehensive FERET evaluation [3, 4, 5, 6], aimed at evaluating different systems using the same, large database containing thousands of images, the systems described in [33, 44, 56, 60, 79], as well as others were evaluated. The neural network method based on Elastic Bunch Graph Matching [79], the statistical method based on subspace LDA [60], and the probabilistic PCA method [33] were adjudged to be among the top three, with each method showing different levels of performance on different subsets of sequestered images. More details on the FERET evaluations will be presented in Section 5.

## 4 Face Recognition from Image Sequences

In surveillance applications, face recognition and identification from a video sequence is an important problem. After over twenty years of research on image sequence analysis [86, 87, 88, 89], only a little of that research had been applied to the face recognition problem [90, 91, 92, 93, 94, 95] up to the mid-nineties. During the last five years, research on human action/behavior recognition from video has been very active. Generic description of human behavior not particular to an individual is interesting and useful. For example, an interactive computer/smart room [96, 97] can recognize such behavior and initiate appropriate action. Another example is the detection of a driver's tiredness [98] by monitoring the driver's facial expressions and head movements. But the task of recognizing individuals from a surveillance video is still difficult for the following reasons:

1. **The quality of the video is low**. Usually videotaping occurs outdoors and the subjects are not cooperative, hence there are possibly large illumination and pose variations in the face images. In addition, partial occlusions and disguise are possible. One possible way to improve the quality of face images is to apply super-resolution techniques [99, 100, 101, 102].

2. **The face images are small.** Again, due to the acquisition conditions, the face image sizes are smaller (sometimes much smaller) than the assumed sizes in most existing still image based face recognition systems. For example, the valid face

region could be as small as $20 \times 20$ pixels, whereas the face image sizes used in still image-based systems are as large as $128 \times 128$. Small-size images not only make the recognition task difficult, but also affect the accuracy of face segmentation, as well as the accurate detection of the fiducial points/landmarks that are often needed in recognition methods.

3. **The characteristics of face/human objects**. One of the main reasons for the feasibility of generic description of human behavior is that the intra-class variation of human objects, and in particular face objects, is much smaller than the objects outside of the class. For the same reason, recognition of individuals within the class is difficult. For example, detecting and localizing faces is much easier than recognizing a specific face.

### 4.1 Basic techniques in video-based face processing

In [9], four computer vision areas were mentioned as being important for video-based face recognition: segmentation of moving objects (humans) from a video sequence; structure estimation; 3-D models for faces; and non-rigid motion analysis. Recent developments in face tracking, modeling and recognition from video verify this prediction. For example, in [103] a face modeling system which estimates facial features and texture from a video stream is described. This system utilizes all four techniques: segmentation of the face based on skin color to initiate tracking; 3D models for the face based on laser-scanned range data to normalize the image (by facial feature alignment and texture mapping into a frontal view) and construct an eigen-subspace for 3D heads; structure from motion at each feature point to provide depth information; and non-rigid motion analysis of the facial features based on simple 2D SSD tracking constrained by a global 3D model. We briefly review these four areas in the following paragraphs.

1. **Video-Based Object Segmentation** Early attempts [44, 104] at segmenting moving faces from an image sequence used simple pixel-based change detection procedures based on difference images. These techniques may run into difficulties when multiple moving objects and occlusion are present. Flow field based methods for segmenting humans in motion are reported in [105]. There is a large body of literature analysis on segmenting/detecting moving objects in video obtained from a stationary or moving platform. Methods based on analysis of difference images, discontinuities in flow fields using clustering, line processes or Markov random field models are available. Some of these techniques have been extended to situations in which both the camera and the objects are moving. A good approach to face segmentation from image sequences is to combine motion detection/clustering and face detection in the individual images. Skin color can also be utilized to enhance the robustness of face detection algorithms.

2. **Structure from Motion** The problem of structure from motion is to estimate the 3-D depths of points from a sequence of images. Unless the camera can be moved along a known baseline, techniques such as motion stereo are not applicable. The structure from motion problem has been approached in two ways. In the differential

approach, one computes some type of flow field (optical, image or normal) and uses it to estimate the depths of visible points. The difficulty in this approach is reliable computation of the flow field. In the discrete approach, a set of features such as points, edges, corners, lines or contours are tracked over a sequence of frames, and the depths of these features are computed. The difficulty here is the correspondence problem — the task of matching the features over a sequence of frames. In both the differential and discrete approaches, the parameters that characterize the motion of the camera appear jointly with the depth parameters. The motion parameters may be useful in predicting where objects will appear in subsequent frames, making the segmentation of these frames somewhat easier. The depth information is useful in building 3-D models for objects and possibly using these models for object recognition in the presence of occlusion. It should be pointed out that if only a monocular image sequence is available, the depth information is available only up to a scale factor; whereas if a binocular (or multi-camera) image sequence is available, one can get absolute depth values using stereo triangulation. Given that laser range finders may not be practical for surveillance applications, multi-camera image sequences may be the best way to get depth information. Another point worth mentioning is that when a discrete approach is used, the depth values are available only at sparse points, requiring interpolation; when a flow-based method is used, dense depth maps can be constructed. Over the last 25 years, hundreds of papers dealing with structure from motion have appeared. It is beyond the scope of this paper to give even a brief summary of major techniques. We list here only books [86, 87, 106, 107, 108, 109] and review papers [110, 111, 112].

3. **3D Models for Faces** 3D models of faces have been employed [113, 114, 115] in the model-based image compression literature by several research groups. Such models are also useful in applications such as forensic face reconstruction from partial information, and computerized aging. They may also be useful for face recognition in the presence of disguises. In [116], real-time 3D modeling and tracking of faces is described; a generic 3D head model is aligned to match frontal views of the face in a video sequence.

4. **Non-rigid Motion Analysis** A final area of relevance to FRT is the motion analysis of non-rigid objects [117, 118, 119, 120]. Some of the work [121, 122] is potentially useful in face recognition. Another application of non-rigid motion to faces to is the recognition of facial expressions from image sequences [123].

### 4.2 Tracking, modeling and non-face-based recognition

During the past five years, tracking, modeling and recognition of hand gestures and human behaviors have been extensively studied. We briefly review some of these topics here. Research on human emotion recognition has been extended to a new area — affective computing [124], in which cues such as facial expressions and body movements, as well as psychological data, are used [125].

### 4.2.1 Tracking and modeling face objects

After faces are located using one of the many methods reviewed earlier, they can be tracked. Face tracking can be divided into three categories: 1) Head tracking, which involves tracking the motion of a rigid object that is performing rotations and translations, 2) Facial feature tracking, which involves tracking non-rigid deformations that are limited by the anatomy of the head, and 3) complete tracking, which involves tracking both the head and the facial features. Early efforts focussed on the first two problems: head tracking [126, 127] and facial feature tracking [121, 122]. In [127], an approach to head tracking using points with high Hessian values was proposed. Several such points on the head are tracked and the 3D motion parameters of the head are recovered by solving an over-constrained set of motion equations. Facial feature tracking methods may make use of the feature boundary or the feature region. Feature boundary tracking attempts to track and accurately delineate the shape of the facial feature, e.g., to track the contours of the lips and mouth [117, 122]. Feature region tracking addresses the simpler problem of tracking a region such as a bounding box that surrounds the facial feature [128]. Facial features are subject to general types of motions: rigid motion due to the head's rotation and translation, articulated motion due to speech or facial expressions, and deformable motion due to muscle contractions and expansions.

In [128], a tracking system based on local parameterized models is used for recognizing facial expressions. The models include a planar model for the head, local affine models for the eyes, and local affine models and curvature for the mouth and eyebrows. A face tracking system was used in [129] to estimate the pose of the face. This system used a graph representation with about 20-40 nodes/landmarks to model the face. Knowledge about faces is used to find the landmarks in the first frame. Two tracking systems described in [103, 116] model faces completely with texture and geometry. Both systems use 3D models and structure from motion to recover the face structure. [103] tracks fixed feature points (eyes, nose tip), while[116] tracks only points with high Hessian values. Also, [103] tracks 2D features in 3D by deforming them, while [116] relies on direct comparison of a 3D model to the image. Methods are proposed in [130, 131] to solve the varying appearance problem in tracking.

An important application of tracking and modeling is to enhance face recognition by providing additional information. After face pose is estimated as in [129], a virtual frontal face can be synthesized, so that the performance of face recognition can be improved. Another useful application of facial feature tracking is the recognition of gaze, based on both head and eye tracking.

### 4.2.2 Recognition of facial expressions

Facial expression recognition has received increased attention during the last five years. Previously, head tracking and facial feature tracking were treated as two separate problems. By jointly addressing the two problems, the recognition of facial expressions has become possible even when large head motion is present.

The rationale for facial expression recognition based on motion can be derived from studies in psychology. Such studies have indicated that at least six emotions are universally associated with distinct facial expressions [132]: happiness, sadness, surprise, fear, anger, and disgust. Most psychological studies of facial expressions have made use of mug-shot images that capture the subject's expression at its peak [133]. Only a few studies have investigated the influence of the motion and deformation of the facial features on the interpretation of facial expressions. Bassili [134] suggested that motion in the image of a face could allow emotions to be identified even with minimal information about the spatial arrangement of the features.

In the engineering literature, early efforts [123, 135] were based on analysis of the optical flow field of the image sequence, which provides clues to the spatial changes in the facial features. [128] demonstrated successful facial expression recognition in extensive laboratory experiments involving 40 subjects as well as in television and movie sequences. 3D motion estimation has also been used to recognize facial expressions [136].

In principle, facial expression recognition can be integrated into a face recognition system so the system is robust to expression variations. In practice, however, it seems that moderate, non-dramatic expressions can be handled by many existing face recognition systems.

### 4.2.3 Recognition of hand gestures

Hand gestures are another important cue to understanding human behavior. They are usually recognized from the temporal characteristics of the hand movements and the poses of the hands during pauses. Hidden Markov Models (HMMs) [137] are the most commonly used tool for gesture recognition. [138] used HMMs to recognize gestures in binary image sequences, using a rotation-invariant representation the images and a neural net. [139] incorporated multiple representations in an HMM framework, using eigenimage weights as features. [140] used geometrical parameters (the image coordinates and orientations of the hands) as image features and employed an HMM five-state topology for gesture classification; good results were reported on classifying 40 American Sign Language gestures in real-time video. In [141], this approach was extended to use 3D measurements obtained from a stereo system as features. Gesture recognition was performed on a set of 18 T'ai Chi gestures (an ancient Chinese martial art), and the performance of ten different feature vectors derived from 3D hand and head tracking data was compared.

Researchers have also done hand sign and pose recognition from still imagery. [142] described a general framework for learning-based hand sign recognition. Discriminant analysis was used to automatically select the most discriminating features and good recognition results were obtained for 28 different static hand signs. [143] applied an elastic graph matching based approach to gesture recognition.

[144] describes the use of hand gesture analysis in combination with speech recognition in a bi-modal interface for controlling a 3D display. For a review of hand gesture recognition techniques, see [145]; for more detailed descriptions of various techniques, see the Proceedings of the AFGR Conferences [1].

### 4.2.4 Recognition of body movement and behavior

Much work has been done on human body tracking. It is impossible to discuss all the relevant references; we only review a few papers briefly here. In [146, 147] the body is modeled by rigid segments that meet at joints. In [148] motion templates are used to track people, in [149] color blobs are used, and in [119] nonrigid models are used. More recently, [150] presents a new visual motion estimation technique that is able to accurately recover high-degree-of-freedom articulated human body configurations in video sequences. This work used a model of the human body consisting of segments attached at joints, subject to constraints involving twist and a product of exponential map.

Pfinder [149] is a real-time person tracking system which uses a multi-class statistical model of color and shape to segment the person from the background. It finds and tracks the person's head and hands under a wide range of viewing conditions. In [151], many levels of representation based on mixture models, EM, recursive Kalman and Markov estimation are used to learn and recognize human dynamics. In [152] a real-time system (the $W^4$ system) for detecting/tracking people and monitoring their activities in an outdoor environment is described. This system operates on monocular gray-scale video imagery or on video imagery from an infrared camera. It uses a combination of shape analysis and robust estimation to locate the people and their parts (head, hands, feet, torso) and to create models of the people's appearances. Building these appearance models enables the system to track the people through occlusions or interactions during which tracking cannot be carried out. The system can also track multiple people through occlusions. In [153] the $W^4$ system was extended to include a stereo matching module. The system is expected to be expanded in the near future to include more modules to recognize various types of actions, e.g. taking leaving, or exchanging objects. In [154], a vision system is described that monitors activities in a site over extended periods of time. The system uses a distributed set of sensors to cover the site, and an adaptive tracker to detect multiple moving objects. The tracker data are used for self-calibration — determining the positions of all the cameras relative to each other; construction of rough site models — determining the ground plane and marking occupied areas; robust detection of objects in the site and classification of the detected objects (e.g., vehicles or pedestrians); and learning common activity patterns, and thereby detecting unusual events in the site from extended observations.

### 4.2.5 Speechreading: enhancing speech recognition

Visual facial cues have been found to be valuable for enhancing speech recognition system performance under noisy conditions [155]. A typical speechreading system consists of two sub-systems: a video sub-system and an audio sub-system. In the video sub-system, a camera captures images of the speaker, which are then digitized and processed to extract useful features for speech recognition. Possible low-level features include the width and height of the mouth, its shape and rounding, the location and velocity of the jaw, and the position of the tongue. Higher-level features include rounding (protrusion of the lips as in /OK/) and the f-tuck (touching of the upper teeth to the lower lip, as in /fa/ and /va/). Important issues in building a speechreading system are how to choose appropriate visual features and how to integrate the video and audio sub-systems. Many papers on this subject have appeared over the past 15 years; examples are [156, 157, 158]. For reviews of this subject see [155, 159].

### 4.3 Video-based face recognition

As mentioned earlier, face recognition in surveillance video is difficult for several reasons. However, there are many situations in which video-based FRT is feasible. For example, in applications such as access control and ATM, the video is acquired in a relatively controlled environment and the face region is also relatively large. In such cases, video-based FRT offers several advantages over still-image-based FRT:

1. Video provides abundant image data; we can select good frames on which to perform classification.

2. Video provides temporal continuity [90]; this allows reuse of classification information obtained from high-quality frames in processing low-quality frames.

3. Video allows tracking of face images; hence phenomena such as facial expressions and pose changes can be compensated for, resulting in improved recognition.

Most video-based FRT systems consist of three modules: a face detection module, a face tracking module, and a face recognition module. Nearly all systems apply still-image-based recognition to selected good frames. The face images are warped into frontal views whenever pose and depth information about the faces is available [95]. Some systems [95] use non-visual cues (speech, for example) to enhance their performance. A number of commercial systems are available — for example, Visionics' FaceIt [160]; however, due to proprietary concerns, their techniques are not open to the public, though their systems may have been initially based on published algorithms.

[90] describes a system for video-based face recognition. This system uses an RBF (Radial Basis Function) network as the learning/recognition engine, and DoG (Difference of Gaussian) filters or Gabor wavelet analysis as the feature representation. The main reasons for the use of a two-layer RBF are its fast learning rate and well-developed mathematical theory. Detection and segmentation of face images is based on motion; the details are not described in [90] and it is indicated that the segmentation results can be imperfect. To train and test the system, two sets of data were used, primary sequences and secondary sequences. Each primary set was a controlled set, including the types of variability the trained system should be tolerant to. Eight primary sequences were collected; each consisted of a person seen against a plain, mid-grey background and turning his head from one profile view to the other. Only one secondary sequence was collected; in it, a person moves from side to side, stopping and starting against a cluttered, changing background. The lengths of the primary sequences were from 62 to 94 frames (a total of 554 frames), while the length of the secondary sequence was 169 frames. Widely varying performance results were reported. For example, only 40% correct classification was obtained when training on 16 frames and testing on the remaining 538 frames from

the primary sequences. On the other hand, 96% correct classification was reported when training on 278 frames and testing on 276 frames, with a 12% rejection rate.

An access control system based on person authentication is described in [91]. The system combines two complementary visual cues: motion and facial appearance. In order to reliably detect significant motions, spatio-temporal zero crossings computed from six consecutive frames were used. These motions were grouped into moving objects using a clustering algorithm and Kalman filters were employed to track the grouped objects. An appearance-based face detection scheme using RBF networks (similar to [38]) was used to confirm that an object is a person. The face detection scheme was "bootstrapped" using motion and object detection to provide an approximate head region. Face tracking based on the RBF network was used to provide feedback to the motion clustering process to help deal with occlusions. Good tracking results were demonstrated, but person authentication results are referred to as future work.

In [92], a fully automatic person authentication system is described which includes video break, face detection, and authentication modules. Video skimming was used to reduce the number of frames to be processed. The video break module, corresponding to key-frame detection based on object motion, consisted of two units. The first unit implemented a simple optical flow method; it was used when the image SNR level is low. When the SNR level was high, simple pair-wise frame differencing was used to detect the moving object. The face detection module consisted of three units: face localization using analysis of projections along the $x$- and $y$-axes; face region labeling using a decision tree learned from positive and negative examples taken from 12 images each consisting of 2759 windows of size $8 \times 8$; and face normalization based on the numbers of face region labels. The normalized face images are then used for authentication, using an RBF network. This system was tested on three image sequences; the first was taken indoors with one subject present, the second was taken outdoors with two subjects, and the third was taken outdoors with one subject in stormy conditions. Perfect results were reported on all three sequences, as verified against a database of 20 still face images.

In [161], a generic approach to simultaneous object tracking and verification is proposed. The approach is based on posterior probability density estimation using sequential Monte Carlo methods [162, 163]. Tracking, which is a temporal correspondence problem, is formulated as a probability density propagation problem, with the density $\pi_t(x)$ being defined over a state space characterizing the object configuration. Using sequential importance sampling, the density is approximated at time $t$ by a set of samples and weights, $\{X_t^{(j)}, w_t^{(j)}\}$. The tracked object is then specified by evaluating the mean value as

$$E_\pi\{X\} \approx \frac{\sum_j X^{(j)} w^{(j)}}{\sum_j w^{(j)}} \tag{13}$$

Using this approach and reparametrization, tracking applications involving different representations such as edge maps, intensity templates, and feature point sets can be uniformly processed by the same algorithm. In addition to tracking, the algoorithm also provides verification results. This is realized by hypothesis testing using the posterior probabilities, which are obtained by integrating (summing, in discrete cases) the esti-

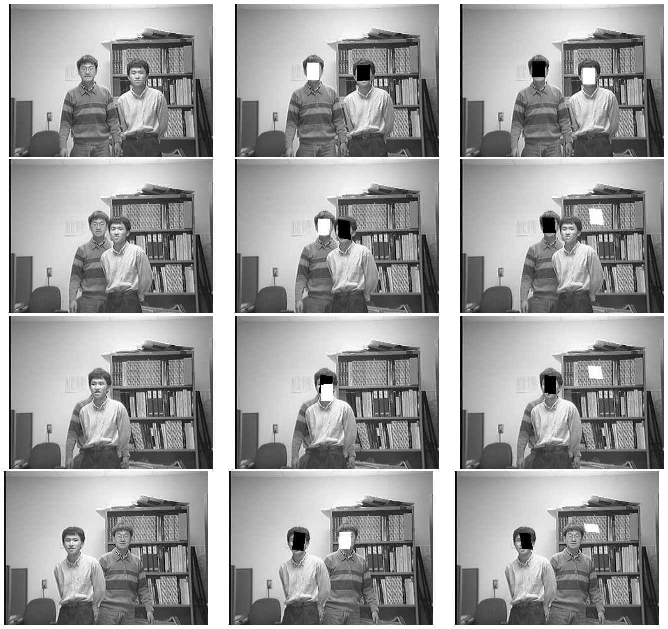


Figure 4: Left: Sample frames of a sequence. Middle: results when correct templates are overlaid on the video; Right: results when the templates are incorrect.

mated densities $\pi_t(x) = p_i(x|Z)$

$$P(\omega_i|Z) = \int_A p_i(x|Z)dx \tag{14}$$

where $\omega_i$ denotes class $i$, $Z$ the observation, and $p_i(x|Z)$ the conditional posterior density for class $i$.

Figure 4 shows sample frames (left column) of a sequence showing two persons moving around; their face templates are used to track and verify them in the video. In the middle and right columns, the templates are overlaid on the video. For easy visualization, a black block is used for the template corresponding to the face of the man in the white shirt (denoted M1), and a white block for the template corresponding to the second man's face (denoted M2). The middle column illustrates the situation where the algorithm is correctly initialized, meaning that the templates are put on the correct persons. The figure shows that tracking is always maintained for M1, and is able to recover from occlusion for M2. The right column shows a case in which the templates were put on the wrong persons. It is seen that M2 eventually drops onto the background, while M1, after tracking the wrong person, is attracted to the right person after they meet. In both cases, the posterior probabilities provide verification results.

The systems described above were tested only on small databases (if at all); their main purpose was to demonstrate the feasibility of video-based face recognition. Two other systems [93, 95] are more practical in terms of accuracy and size of the database. Both of these systems use more than one cue; for example [95] uses both audio and video, and [93] uses stereo. (For more information about recognition based on video and audio see the Proceedings of the AVBPA Conferences [2].)

In [93], a system called PersonSpotter is described. This system is able to capture, track and recognize a person walking toward or passing a stereo CCD camera. It has several modules, including a head tracker, preselector, landmark finder, and identifier. The head tracker determines the image regions that are changing due to object motion based on simple image differences. A stereo algorithm then determines the stereo disparities of these moving pixels. These disparity values are used to compute histograms for image regions. Regions with in a certain disparity interval are selected and referred to as "silhouettes". Two types of detectors, skin color based and convex region based, are applied to these silhouette images. The outputs of these detectors are clustered to form regions of interest which usually correspond to heads. To track a head robustly, temporal continuity is exploited in the form of the thresholds used to initiate, track and delete an object.

To find the face region in an image, the preselector uses a generic sparse graph consisting of 16 nodes learned from 8 example face images. The landmark finder uses a dense graph consisting of 48 nodes learned from 25 example images to find landmarks such as the eyes and the nose tip. Finally, an elastic graph matching scheme is employed to identify the face. For details about these modules, see [93]. A recognition rate of about 90% was achieved (the size of the database is not known). The system was implemented using ANSI C++ on a Unix platform (a four-processor 90-Mhz SGI) and was able to process 6-8 persons per minute. The size of the normalized face images should be about the same as [77], i.e. about 128 × 128 pixels.

A multimodal based person recognition system is described in [95]. This system consists of a face recognition module, a speaker identification module, and a classifier fusion module. It has the following characteristics:

- The face recognition module can detect and compensate for pose variations; the speaker identification module can detect and compensate for changes in the auditory background.
- The most reliable video frames and audio clips are selected for recognition.
- 3D information about the head is used to detect the presence of an actual person as opposed to an image of that person.

Recognition and verification rates of 100% were achieved; for 26 registered clients.

The face recognition module consists of three units:

(a) **Face Detection and Tracking** The face is detected using skin color information using a learned model of a mixture of Gaussians. The facial features (eyes, mouth, nose) are then located using symmetry transforms and image intensity gradients. Correlation-based methods are used to track the feature points. The locations of these feature points are used to estimate the pose of the face. This pose estimate and a 3D head model are used to warp the detected face image into a frontal view.

(b) **Eigenspace Modeling** For recognition, the feature locations are refined and the face is normalized with eyes and mouth in fixed locations. Images from the face tracker are used to train a frontal eigenspace, and the leading 35 eigenvectors are retained. Since the face images have been warped into frontal views a single eigenspace is enough. Face recognition is then performed using the eigenface approach with additional temporal information added. The projection coefficients of all images of each person are modeled as a Gaussian distribution, and the face is classified based on the probability of match.

(c) **Depth Estimation** For greater robustness, depth information at each feature position is obtained using an SfM technique. This depth information can be used to distinguish a real head from a head image; this makes it difficult to fool the system with still face images.

The speaker recognition module has four units:

(a) **Event Detection** Coarse segmentation of the audio is used to identify segments that are likely to contain speech. This segmentation is performed using a simple event detector constructed by thresholding the total energy and incorporating constraints on event length and surrounding pauses.

(b) **Feature Extraction** The (16kHz sampled) audio is then filtered with a weak high-pass filter to remove DC offsets and boost higher frequencies. Mel-scaled frequency coefficients (MFCs) are then computed for audio frames 32 ms long and 16 ms apart.

(c) **Modeling** An HMM, estimated from speech samples, is used to model the spectral signature of each person's speech. However, experiments showed that 1-state HMM models with 30 Gaussians performed best, suggesting that the use of HMMs is unnecessary.

(d) **Background Adaption** It is well known that statistical models trained on clean speech perform poorly in noisy or altered environments. Two common types of noise are convolutional noise (due primarily to the use of different microphones and sound cards) and additive noise (due to the presence of other sound sources). Here only additive noise was considered, and HMMs were used to model the clean speech, the additive noise, and the combination of both.

Finally, the face and speaker recognition modules are combined using a Bayes net. The system was tested in an ATM scenario. An ATM session begins when the subject enters the camera's field of view and the system detects his/her face. The system then greets the user and begins the banking transaction, which involves a series of questions by the system and answers by the user. Data for 26 people were collected; the normalized face images were $40 \times 80$ pixels and the audio was sampled at 16 kHz. The experiments showed that the combination of audio and video improved performance, and that perfect (100%) recognition and verification were achieved when the image/audio clips with highest confidence scores were used.

## 5 Evaluation of Face Recognition Systems

Given the numerous theories and techniques that are applicable to face recognition, it is clear that evaluation and benchmarking of these algorithms is crucial. Previous work on the evaluation of OCR and fingerprint classification systems [164, 165] provided insights into how the evaluation of algorithms and systems can be performed efficiently. One of the most important facts learned in these evaluations is that large sets of test images are essential for adequate evaluation. It is also extremely important that the sample be statistically as similar as possible to the images that arise in the application being considered. Scoring should be done in a way that reflects the costs of errors in recognition. Reject-error behavior should be studied, not just forced recognition.

In planning an evaluation, it is important to keep in mind that the operation of a pattern recognition system is statistical, with measurable distributions of success and failure. These distributions are very application-dependent, and no theory seems to exist that can predict them for new applications. This strongly suggests that an evaluation should be based as closely as possible on a specific application.

During the past five years, several large, publicly available face databases have been collected and corresponding testing protocols have been designed. The series of FERET evaluations [3, 4, 5, 6] attracted many institutions and companies to participate. We describe here the two most important face databases and their associated evaluation methods.

### 5.1 The FERET protocol

Until recently, there did not exist a common FRT evaluation protocol that included large databases and standard evaluation methods. This made it difficult to assess the status of FRT for real applications, even though many existing systems reported almost perfect performance on small databases. Measuring the performance of FRT in a framework that models real-world settings was one of the three primary goals of the FERET program [166, 167, 168, 5]. The other two goals were advancing FRT and collecting a large database of facial images to support algorithm development and evaluation. The database was collected between August 1993 and July 1996, and consists of 14,126 images of 1199 individuals. The FERET database was divided into development and sequestered portions. The development portion was made available to researchers for algorithm development, and the sequestered portion was retained for independent evaluation and testing of algorithms. In late 2000 the entire FERET database is being released along with the Sep96 evaluation protocols, evaluation scoring code, and baseline PCA algorithms.

The first FERET evaluation test (Aug94) was administered in August 1994 [168]. This evaluation established a baseline for face recognition algorithms, and was designed to measure performance of algorithms that could automatically locate, normalize, and identify faces. This evaluation consisted of three tests, each with a different gallery and probe set. (A gallery is a set of known individuals, while a probe is a set of unknown faces presented to a system for recognition.) The first test measured identification performance from a gallery of 316 individuals with one image per person; the second was a false-alarm test, and the third measured the effects of pose changes on performance. The second

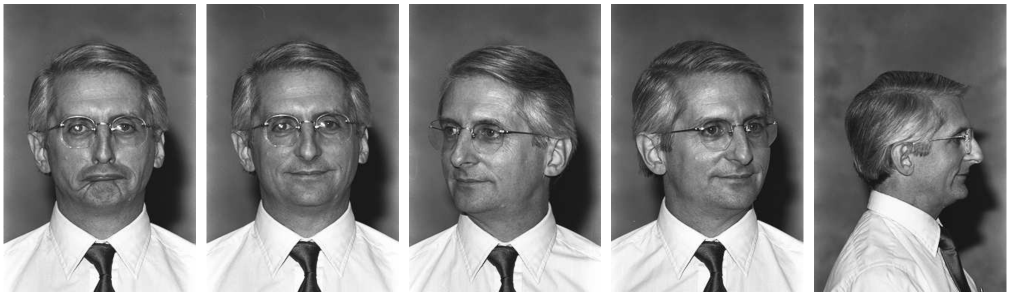


Figure 5: Images from the FERET dataset; these images are of size 384 × 256.

FERET evaluation (Mar95) was administered in March 1995; it consisted of a single test that measured identification performance from a gallery of 817 individuals, and included 463 duplicates in the probe set [168]. (A duplicate is a probe for which the corresponding gallery image was taken on a different day; there were only 60 duplicates in the Aug94 evaluation.) The third and last evaluation (Sep96) was administered in September 1996 and March 1997. The design of this evaluation was more complex than the first two evaluation, and allowed for more detailed performance characterization of face recognition systems.

### 5.1.1 Database

Currently, the FERET database is the only large database that is generally available to researchers without charge [167, 168]. The images in the database were initially acquired with a 35-mm camera, using color Kodak Ultra film. The film was processed by Kodak and stored on CD-ROM using Kodak's multiresolution technique for digitizing and storing imagery. The color images were retrieved from CD-ROM and converted into 8-bit gray scale images. Each image was assigned a file name that encoded the subject's identity, the date the image was taken, the nominal pose of the head, and special variations.

The images were collected in 15 sessions between August 1993 and July 1996. Each session lasted one or two days, and the location and setup did not change during the session. Sets of 5 to 11 images of each individual were acquired under relatively unconstrained conditions; see Figure 5. They included two frontal views; in the first of these (**fa**) a neutral facial expression was requested and in the second (**fb**) a different facial expression was requested (these requests were not always honored). For 200 individuals, a third frontal view was taken using a different camera and different lighting; this is referred to as the **fc** image. The remaining images were non-frontal and included right and left profile, right and left quarter profile, and right and left half profile. The FERET database consists of 1564 sets of images (1199 original sets and 365 duplicate sets) — a total of 14,126 images. A development set of 503 sets of images were released to researchers; the remaining images were sequestered by the Government for independent evaluation.

### 5.1.2 Evaluation

For details of the three FERET evaluations see [166, 167, 168, 5]. The results of the Sep96 FERET evaluation, the most recent, will be briefly reviewed here. This evaluation was administered in September 1996 and March 1997. Each algorithm was given two sets of images, a target set and a query set; these are different from the gallery and probe sets used to compute performance statistics. The target set contained 3323 images and the query set contained 3816 images. An algorithm reported a similarity score for all pairs of images taken, respectively, from the target and query sets; this resulted in 12,680,568 similarity scores. (A similarity score is an estimate of how similar two faces are.) Because of the design of the target and query sets, different galleries were constructed from the target set, and different probe sets were constructed from the query set. This allowed for more comprehensive reporting of performance statistics for a larger range of conditions than the first two evaluations. For the Sep96 evaluation, there was a primary gallery consisting of one frontal image (**FA**) per person for 1196 individuals. This was the core gallery used to measure performance for the following four different probe sets.

- **FB** probes—Gallery and probe images of an individual taken on the same day with the same lighting (1195 probes).

- **fc** probes—Gallery and probe images of an individual taken on the same day with different lighting (194 probes).

- Dup I probes—Gallery and probe images of an individual taken on different days—duplicate images (722 probes).

- Dup II probes—Gallery and probe images of an individual taken over a year apart (the gallery consisted of 894 images; 234 probes).

Performance was measured using two basic methods. The first measured identification performance, where the primary performance statistic is the percentage of probes that are correctly identified by the algorithm. The second measured verification performance, where the primary performance measure is the equal error rate between probability of false alarm and probability of correct verification. (A more complete method of reporting identification performance is a cumulative match characteristic; for verification performance it is a receiver operating characteristic (ROC).)

The Sep96 evaluation tested the following ten algorithms:

- An algorithm from Excalibur Corporation (Carlsbad, CA)(Sept. 1996)

- Two algorithms from MIT Media Laboratory (Sept. 1996) [44, 169]

- Three Linear Discriminant Analysis based algorithms from Michigan State University [56] (Sept. 1996) and the University of Maryland [59, 60] (Sept. 1996 and March 1997)

- A gray-scale projection algorithm from Rutgers University [170] (Sept. 1996)

- An Elastic Graph Matching algorithm from the University of Southern California [79, 171] (March 1997)
- A baseline PCA algorithm [44, 172, 173]
- A baseline normalized correlation matching algorithm.

The month the evaluation was administered is given in parentheses.

Three of the algorithms performed very well: Probabilistic Eigenface from MIT [169], Subspace LDA from UMD [60, 63], and Elastic Graph Matching from USC [79]. To separate recognition from localization, two versions of the evaluation were designed: A fully automatic version in which the facial features needed to located, and a semi-automatic version in which the eye coordinates were given. All of the above algorithms took the semi-automatic version, and one of the MIT algorithms and the USC algorithm took the fully automatic version. Noticeable (but not significant, especially for the USC algorithm) performance degradation was observed.

A number of lessons were learned from the FERET evaluations. The first is that performance depends on the probe category and there is a difference between best and average algorithm performance. This is shown in Figure 6, which plots the best and average performance of the partially automatic algorithms from the Sep96 evaluation. This is especially true for the **fc** probes (lighting change). The second is that the availability of a large data set combined with periodic evaluations led to measurable increases in performance. This is directly supported by the results for the MIT and UMD algorithms. In September 1996 two MIT algorithms were tested. One of them was the same algorithm that was tested in March 1995, and the second was a new algorithm developed since March 1995. For UMD, one algorithm was tested in September 1996, and a second algorithm was tested in March 1997. Results from the September 1996 evaluation were used as input in designing the second algorithm. The identification rates for the **FB** and Dup I probe categories are shown in Figures 7 and 8. A third lesson is that the scenario has an impact on performance. For identification, on the **FB** and duplicate probes, the USC scores were 94% and 59%, and the UMD scores were 96% and 47%. However, for verification, the equal error rates were 2% and 14% for USC, and 1% and 12% for UMD. The verification equal error rates for the **FB** and Dup I probe categories are shown in Figures 9 and 10.

Detailed and robust testing can provide insight into the underlying properties of algorithms. During the period of the FERET evaluations there was a debate in the face recognition community about what was the best representation for faces; in fact, the debate is ongoing. A large number of the representations were projection-based. In this class of representations, an $N \times M$ image is interpreted as a point in $N \times M$-dimensional Euclidean space, and the algorithm represents a face in a linear subspace of much lower dimension. The mapping from the original image space to the subspace is a linear projection. Faces are identified by a nearest neighbor classifier. Two examples are PCA and LDA-based algorithms.

The FERET evaluations compared a number of these competing representations; however, the evaluations did not systematically compare different implementations of the same representation. Moon and Phillips [172, 173] systematically compared different

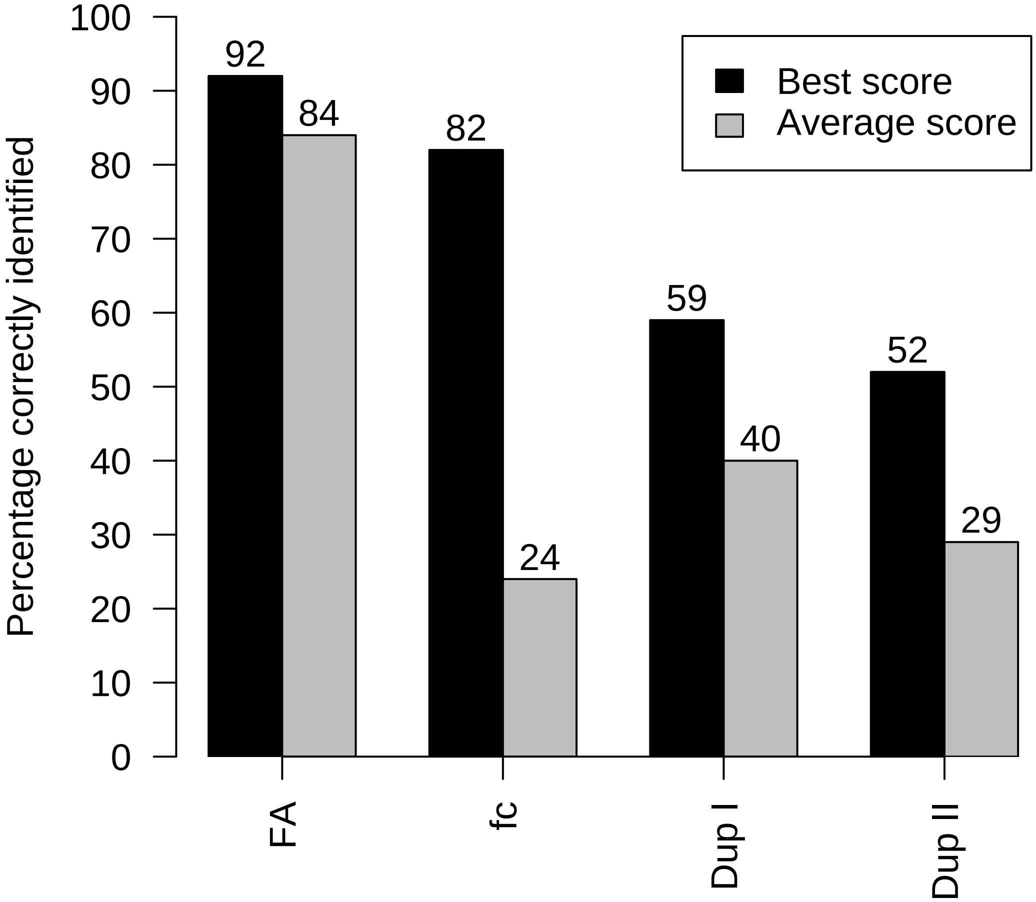


Figure 6: The best and average identification scores for the FERET Sep96 evaluation by probe category. The probe categories are: **FB**, **fc**, Dup I, and Dup II.

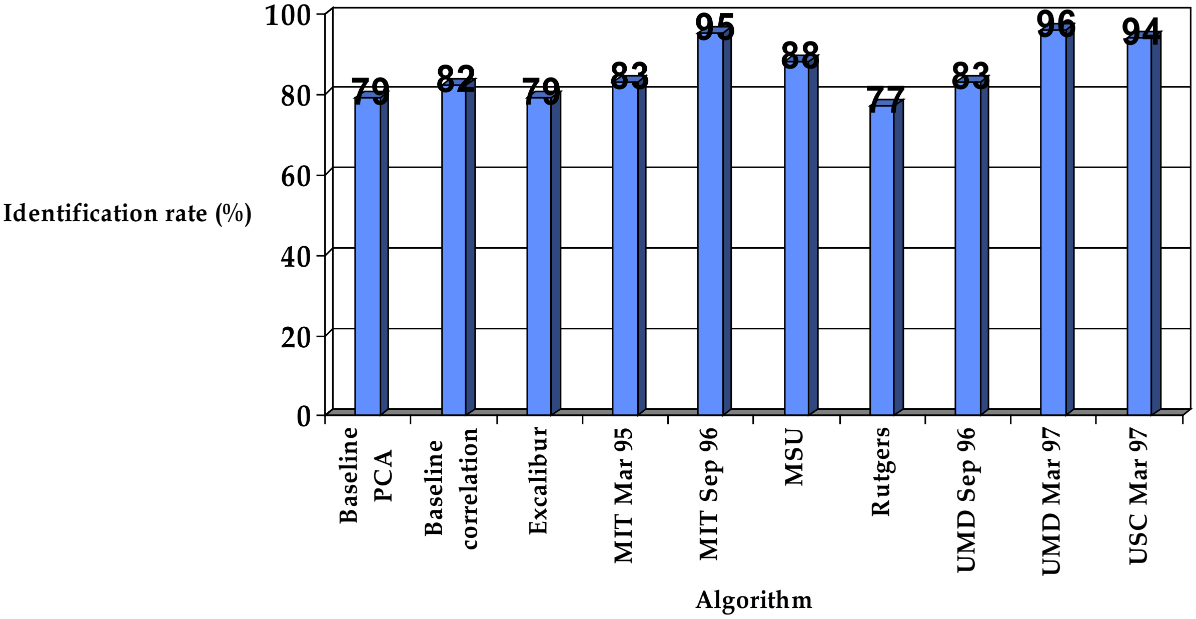


Figure 7: Identification rate for **FB** probes (gallery: 1196, probe: 1195).

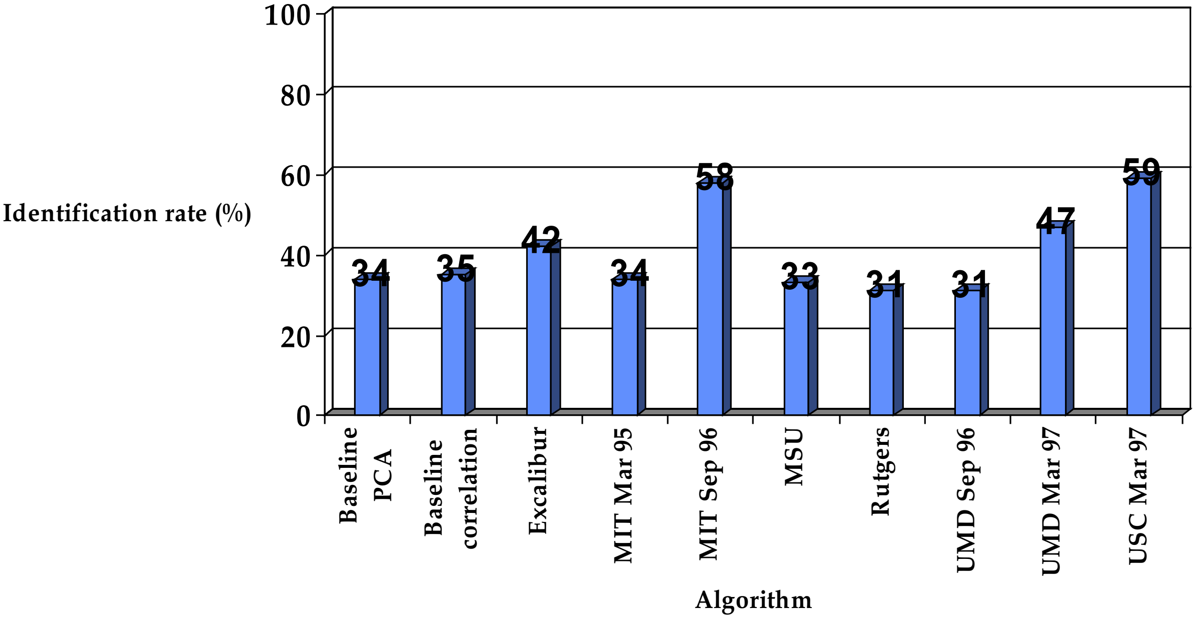


Figure 8: Identification rate for DupI probes (gallery: 1196, probe: 722).

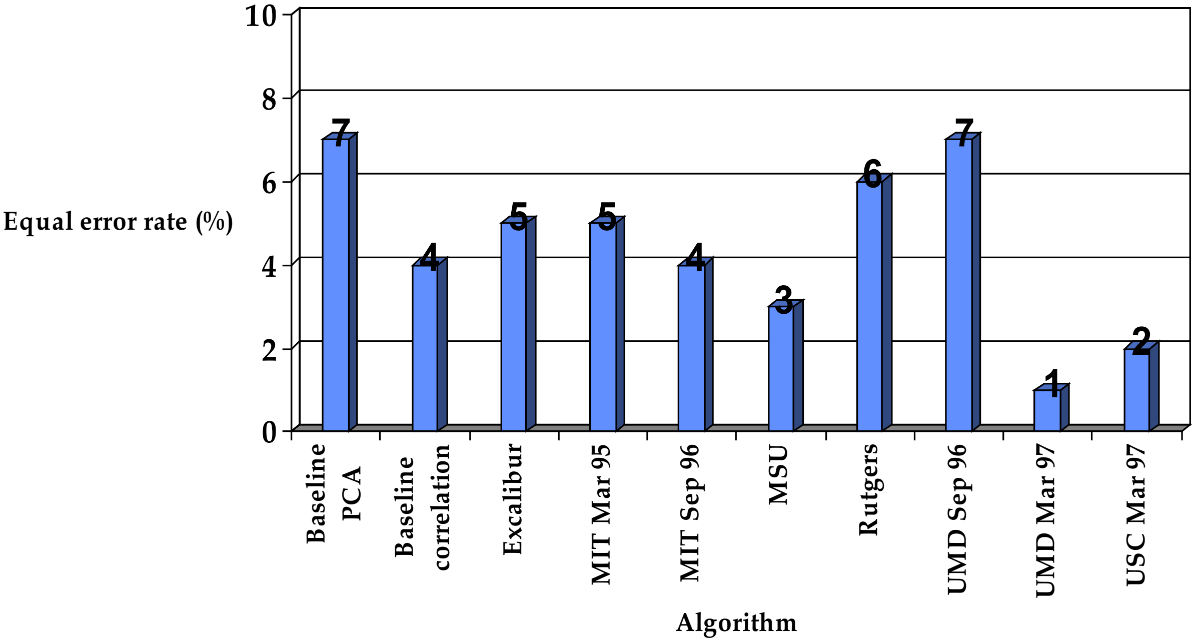


Figure 9: Equal error rate for **FB** probes (gallery: 1196, probe: 1195).

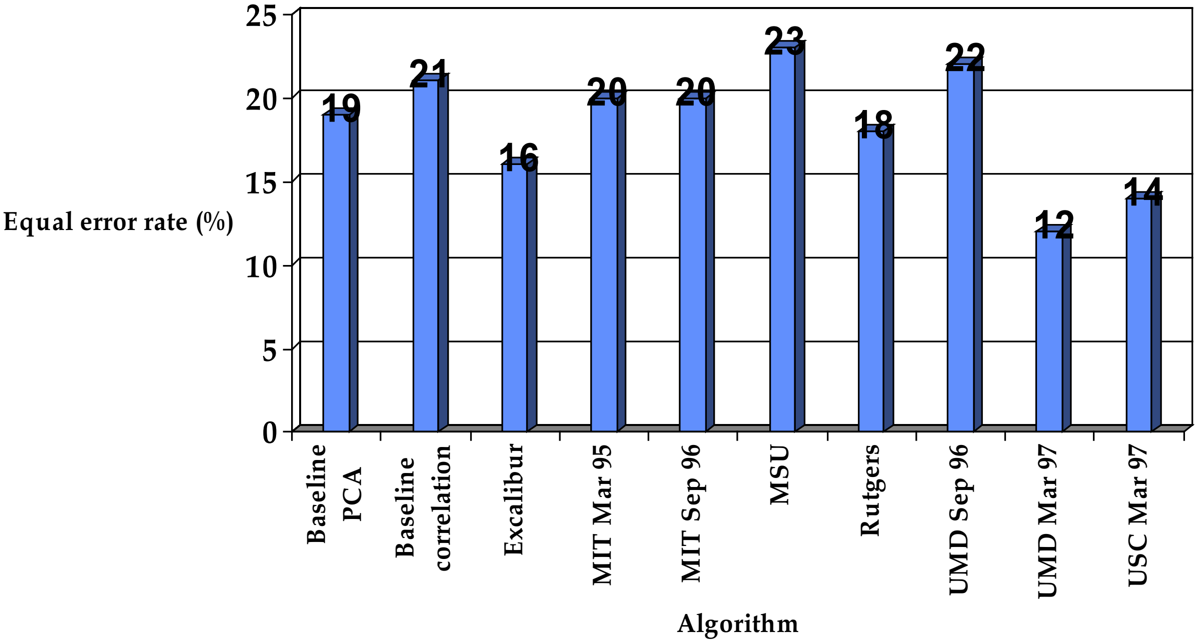


Figure 10: Equal error rate for DupI probes (gallery: 1196, probe: 730).

implementations of a PCA-based face recognition algorithm. For each implementation, the Sep96 performance scores were computed. One class of variations examined was the use of seven different distance metrics in the nearest neighbor classifier, which they found was the most critical element in their implementation. On the Dup I probes, the range of performance for the seven different classifiers is roughly the same for the projection-based algorithms evaluated in the Sep96 evaluation; see Figure 11. This raises the question of what is a more important factor in algorithm performance, the representation or the specifics of the implementation. It also shows the importance of an accepted evaluation methodology and a detailed scientific investigation into the different aspects of an implementation.

**MPEG-7 Evaluation** The subspace LDA algorithm has also been tested on an MPEG-7 content set. In [70], a proposal entitled Descriptor for Human Face Image Objects in Multimedia Databases was submitted to MPEG-7 using the subspace LDA method. The performance of this proposed descriptor in retrieving face image objects from a database was evaluated using MPEG-7 Test Content Set S4 [69]. This set contains a total of 178 face images obtained from 14 different persons (classes). Of the 178 images, 140 are frontal views and the rest are non-frontal views (rotated out of the image plane). The querying procedure usually consisted of two steps: processing of the input image to obtain a representation; and ranking of the retrieved items with respect to a similarity measure.

Subspace LDA projection coefficients were proposed as descriptors of a face image object. The full representation for a face database is the PCA projection matrix $\Phi$, the LDA projection matrix $W$, and the vector $\mathbf{z}$ for each face image. The PCA projection $\Phi$ (of dimension $2016 \times 300$) was computed from the 1038 FERET images, and the LDA projection $W$ was computed from the available MPEG-7 images, which consisted of 14 classes, yielding a matrix $W$ of size $300 \times 13$. Five images selected from each of the 14 classes (one of them a non-frontal view) were used to compute the matrix $W$. Two experiments were performed:

- **Full querying** All 70 images used in the LDA training stage were stored in the database. Each of the 178 available images was used as a query image, and retrieval from the database was performed using subspace LDA. Using the criterion that the top-ranked retrieved image must belong to the correct class, a correct retrieval rate of 86.5% was obtained. Using the criterion that one of the three top-ranked retrieved images must belong to the correct class, a correct retrieval rate of 90.4% was obtained.

- **Frontal-view querying** All 70 training images were again stored in the database, and each of the 130 available frontal-view images was used as a query image. Correct retrieval rates of 93.1% and 95.4% were obtained using the top-ranked and three-top-ranked criteria, respectively.

Some of the query images and the corresponding three top-ranked retrieved database images are shown in Fig. 12.

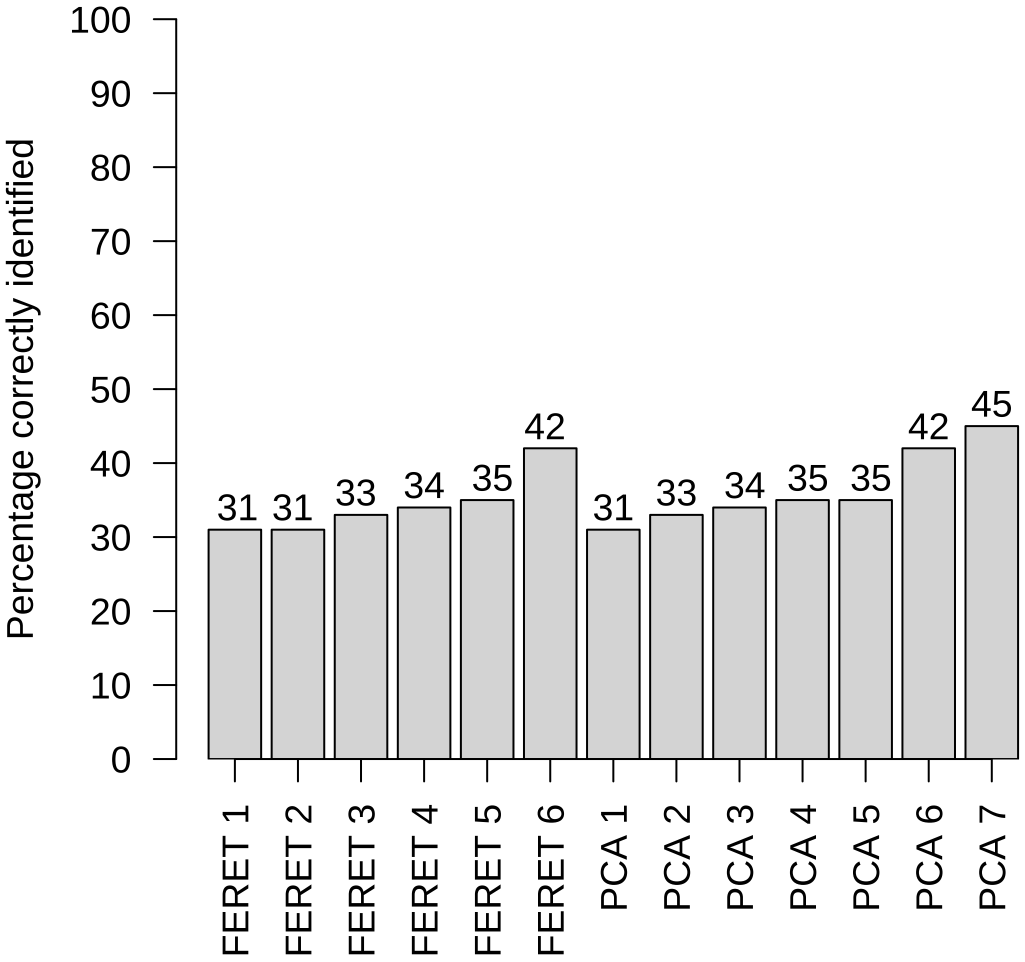
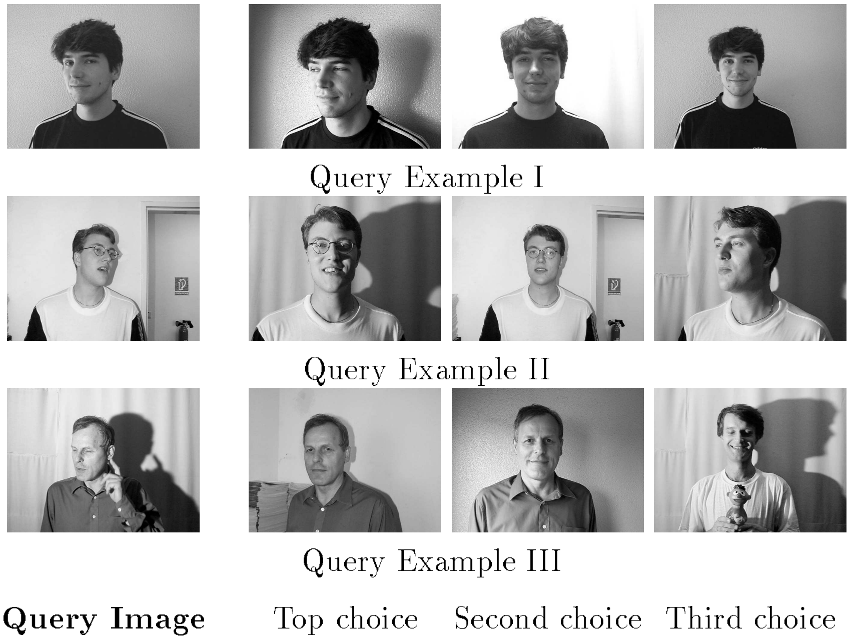


Figure 11: Comparison of six projection-based algorithm that took the Sep96 evaluation and seven different implementations of a PCA-based algorithm. The six projection-based algorithms are labeled FERET 1 through FERET 6, and the seven PCA-based algorithms are labeled PCA 1 through PCA 7. Identification results for the Dup I probes are presented.



Query Example I

Query Example II

Query Example III

Figure 12: Query examples for the MPEG-7 image database.

#### 5.1.3 Summary

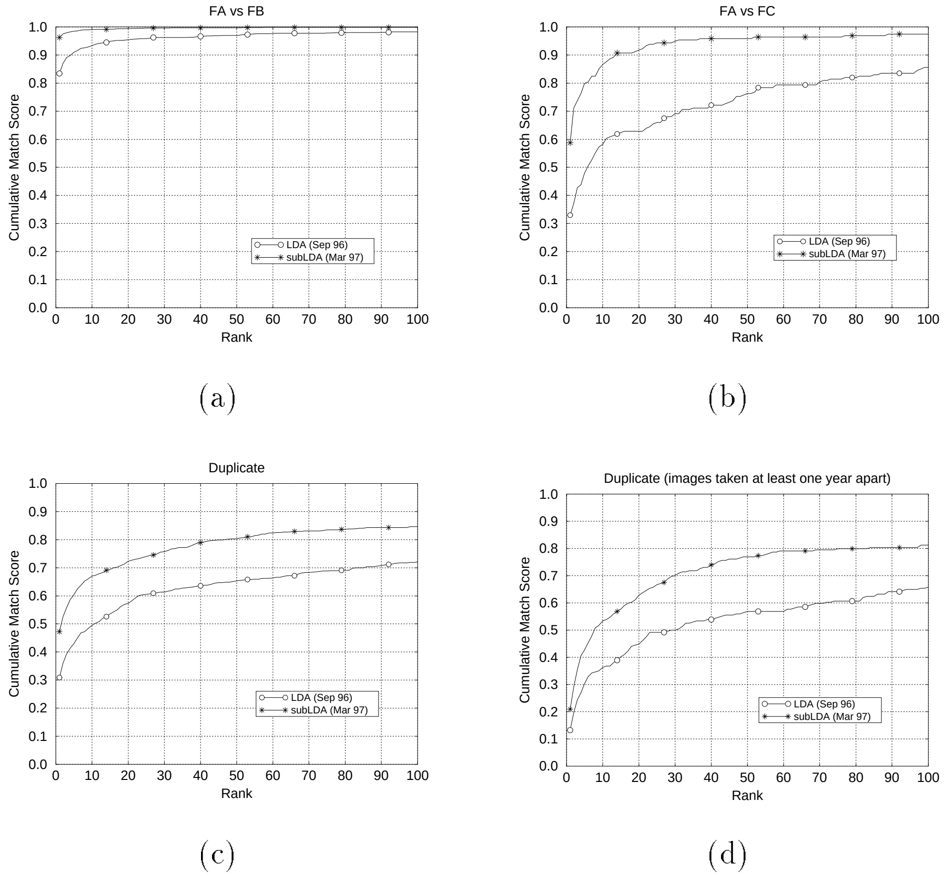
The availability of the FERET database and evaluation technology has had a significant impact on progress in the development of face recognition algorithms. Because of its use of a large database and independent tests, the FERET program has made it possible to objectively evaluate algorithms under close to real-world conditions. The series of tests has allowed advances in algorithm development be quantified — for example, the performance improvements in the MIT algorithms between March 1995 and September 1996, and in the UMD algorithms between September 1996 and March 1997 (Fig. 13).

Another important contribution of the FERET program is the identification of areas for future research. In particular, the August 1994 test suggested two directions for future research: recognition from images collected months or years apart, and recognition under pose changes.

The March 1995 test measured the performance of the algorithms on a larger database that contained more duplicates. Absolute performances on the 1994 and 1995 tests were comparable, in spite of the increase in difficulty, indicating that steady advances were being made in face recognition capability.

Based on the previous tests, an important goal of the September 1996 test was to study the ability of algorithms to recognize people from images taken days, months, or years apart. In general the test results revealed two major problem areas: recognizing duplicates and recognizing people under illumination variations.

The FERET evaluation protocol is the basis of the Face Recognition Vendor Test 2000 and the HumanID database.



(a)

(b)

(c)

(d)

Figure 13: UMD FERET test results from September 96 and March 97: (a)**fa** vs **fb**, (b)**fa** vs **fc**, (c)Duplicate, (d)Duplicate (images taken at least one year apart). [Courtesy of Army Research Laboratory]

### 5.2 The XM2VTS protocol

Multi-modal methods are a very promising approach to user-friendly (hence acceptable), highly secure personal verification. Recognition and verification systems need training; the larger the training set, the better the performance achieved. The volume of data required for training a multi-modal system based on analysis of video and audio signals is on the order of TBytes; technology that allows manipulation and effective use of such volumes of data has only recently become available in the form of digital video. The XM2VTS multimodal database [7] contains four recordings of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot. Available data from this database include high-quality color images, 32 KHz 16-bit sound files, video sequences, and a 3D model.

The XM2VTS database is an expansion of the earlier M2VTS database [174]. The M2VTS project (Multi-Modal Verification for Teleservices and Security Applications), a European ACTS (Advanced Communications Technologies and Services) project, deals with access control by multimodal identification of human faces. The goal of the project was to improve recognition performance by combining the modalities of face and voice. The M2VTS database contained five shots of each of 37 subjects. During each shot, the subjects were asked to count from '0' to '9' in their native language (most of the subjects were French speaking) and rotate their heads from $0°$ to $-90°$, back to $0°$, and then to $+90°$. They were then asked to rotate their heads again with their glasses off, if they wore any. Three subsequences were extracted from these video sequences: voice sequences, motion sequences, and glasses-off motion sequences. The voice sequences can be used for speech verification, frontal view face recognition, and speech/lips correlation analysis. The other two sequences are intended for face recognition only.

It was found that the subjects were relatively difficult to recognize in the fifth shot because it varied significantly in face/voice/camera setup from the other shots. Several experiments have been conducted using the first four shots [175, 176, 177, 178, 179, 180], with the goals of investigating

- text-dependent speaker verification from speech
- text-independent speaker verification from speech
- facial feature extraction and tracking from moving images
- verification from an overall frontal view
- verification from lip shape
- verification from depth information (obtained using structured light)
- verification from a profile
- synchronization of speech and lip movement

### 5.2.1 Database

The XM2VTS database differed from the M2VTS database primarily in the number of subjects (295 rather than 37). The M2VTS database contained five shots of each subject taken at sessions over a period of three months; the XM2VTS database contained eight shots of each subject taken at four sessions over a period of four months (so that each session contains two repetitions of the sequence).

The XM2VTS database was acquired using a Sony VX1000E digital camcorder and a DHR1000UX digital VCR. This VCR captures video at a color sampling resolution of 4:2:0 and audio at a frequency of 32kHz and a sampling rate of 16 bits. It was chosen because it can be interfaced to a computer via a firewire port. At present only the PC architecture is supported, but SUN, SGI and DEC are all working on firewire solutions. Software has been written that allows a user to move video and audio sequences and individual frames to a PC's hard disk directly from the VCR.

In the XM2VTS database, the first shot is a speaking head shot. Each subject, who wore a clip-on microphone, was asked to read three sentences that were written on a board positioned just below the camera. The subjects were asked to read the three sentences twice at their normal pace and to pause briefly at the end of each sentence. The three sentences, which were the same in all four recording sessions, were

1. "0 1 2 3 4 5 6 7 8 9"

2. "5 0 6 9 2 8 1 3 7 4"

3. "Joe took father's green shoe bench out"

The audio sentences were stored in 7080 files which are available on four CDROMs as mono, 16BIT, 32 KHz, and PCM wave files.

The second shot is a rotating head sequence. Each subject was asked to rotate his/her head to the left, to the right, up, and down, and finally to return to the center. The subjects were told that a full profile was required and were asked to repeat the entire sequence twice. The same sequence was used in all four sessions. A set of profile images is available on four CDROMs. These consist of one left profile and one right profile of each person from each session — a total of 2,360 images. The images are stored in color PPM format at a resolution of 720 × 576. A set of frontal images, one per subject per session (1,180 images in all), is also available on two CDROMs.

An additional dataset containing a 3D model of each subject's head was acquired during each session using a high-precision stereo-based 3D camera developed by the Turing Institute[1]. This data set is available in the form of 295 VRML models and texture images on a single CDROM.

### 5.2.2 Evaluation

A protocol was designed (see Tables 2 and 3) to evaluate the performance of vision- and speech-based person authentication systems on the XM2VTS database. This protocol

[1] Turing Institute Web Address: http://www.turing.gla.ac.uk/

was defined for the task of verification. The features of the observed person are compared with stored features corresponding to the claimed identity, and the system decides whether the identity claim is true or false on the basis of a similarity score. The subjects whose features are stored in the system's database are called *clients*, whereas persons claiming false identity are called *imposters*.

The database is divided into three parts: a training set, an evaluation set, and a test set. The training set is used to build client models. The evaluation set is used to compute client and imposter scores. On the basis of these scores, a threshold is chosen that determines whether a person is accepted or rejected. In multi-modal classification, the evaluation set can also be used to optimally combine the outputs of several classifiers. The test set is selected to simulate a real authentication scenario. 295 subjects were randomly divided into 200 clients, 25 evaluation imposters and 70 test imposters. Two different evaluation configurations were used (see Tables 2 and 3) with different distributions of client training and client evaluation data.

| Session | Shot | Clients | Imposters | |
|---|---|---|---|---|
| 1 | 1 | Training | Evaluation | Test |
| | 2 | Evaluation | | |
| 2 | 1 | Training | | |
| | 2 | Evaluation | | |
| 3 | 1 | Training | | |
| | 2 | Evaluation | | |
| 4 | 1 | Test | | |
| | 2 | | | |

Table 2: Configuration I of the evaluation protocol

| Session | Shot | Clients | Imposters | |
|---|---|---|---|---|
| 1 | 1 | Training | Evaluation | Test |
| | 2 | | | |
| 2 | 1 | | | |
| | 2 | | | |
| 3 | 1 | Evaluation | | |
| | 2 | | | |
| 4 | 1 | Test | | |
| | 2 | | | |

Table 3: Configuration II of the evaluation protocol

### 5.2.3 Summary

The results of the M2VTS/XM2VTS projects can be used for a broad range of applications. In the telecommunication field, the results should have a direct impact on network services where security of information and access will become increasingly important. (Telephone fraud in the U.S. has been estimated to cost several billion dollars a year.)

## 6 Two Challenges: Illumination and Pose Variation

Though many face recognition techniques have been proposed and have demonstrated significant promise, the task of robust face recognition is still difficult [68]. The recent FERET test revealed that there are at least two major challenges: The illumination variation problem and the pose variation problem. Either of these problems may cause serious performance degradation for most existing systems. For example, change in illumination conditions can change the 2D appearance (face image) of a 3D face object dramatically, and hence can seriously affect system performance. These two problems have been documented in many evaluations of FRT systems [4, 181] and in the divided opinions of the psychology community [24, 23, 15]. Unfortunately, they are unavoidable when face images are acquired in an uncontrolled environment as in surveillance video clips. In this section, we examine the two problems and review some approaches to solving them. We also point out the pros and cons of these approaches so an appropriate approach can be applied to a specific task.

### 6.1 The illumination problem in face recognition

The illumination problem is illustrated in Fig. 14 where the same face appears different due to a change in lighting. The changes induced by illumination are often larger than the differences between individuals, causing systems based on comparing images to misclassify input images. This was experimentally observed in [181] using a dataset of 25 individuals, and was theoretically proved in [182] for systems based on eigenface projection.

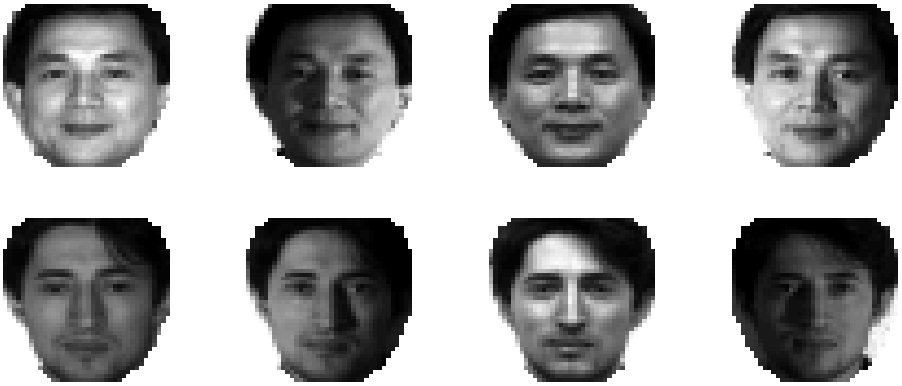


Figure 14: The same face appears differently under different illuminations.

The illumination problem is quite difficult and has received consistent attention in the image understanding literature. In the case of face recognition, many approaches to this problem have been proposed that make use of domain knowledge, in particular of the knowledge that all faces belong to one face class. These approaches can be divided into four types [68]: 1) heuristic methods, e.g. discarding the leading principal components, 2) image comparison methods in which appropriate image representations and distance measures are used, 3) class-based methods using multiple images of the same face in a fixed pose but under different lighting conditions, and 4) model-based approaches in which 3D models are employed.

### 6.1.1 Heuristic Approaches

When the face eigen-subspace domain is used, it has been suggested that by discarding the three most significant principal components, variations due to lighting can be reduced. It was experimentally verified in [58] that discarding the first few principal components works reasonably well for images obtained under different lighting conditions. However, in order to maintain system performance for normally illuminated images, while improving performance for images acquired under changes in illumination, it must be assumed that the first three principal components capture only variations due to lighting. In [52], a heuristic method based on face symmetry was proposed to enhance system performance under lighting changes.

### 6.1.2 Image Comparison Approaches

In [181], approaches based on image comparison using different image representations and distance measures were evaluated. The image representations used were edge maps, derivatives of the gray level, images filtered with 2D Gabor-like functions, and a representation that combines a log function of the intensity with these representations. The distance measures used were pointwise distance, regional distance, affine-GL (gray level) distance, local affine-GL distance, and log pointwise distance. For more details about these methods and about the evaluation database, see [181]. It was concluded that none of these representations alone can overcome the image variations due to illumination. A recently proposed image comparison method [183] used a new measure robust to illumination change. This method is based on the observation that the difference between two images of the same object is smaller than the difference between images of different objects. However, the proposed measure is not strictly illumination-invariant.

### 6.1.3 Class-Based Approaches

Under the assumptions of Lambertian surfaces and no shadowing, a 3D linear illumination subspace for a person was constructed in [66, 67, 184, 185] for a fixed viewpoint, using three aligned faces/images acquired under different lighting conditions. Under ideal assumptions, recognition based on this subspace is illumination-invariant. More recently, an illumination cone has been proposed as an effective method of handling illumination variations, including shadowing and multiple light sources [185, 186]. This method is an extension of the 3D linear subspace method [66, 67] and also requires three aligned training images acquired under different lighting conditions. One drawback of using this method is that more than three aligned images per person are needed. More recently, a method based on a quotient image was introduced [187]. An advantage of this approach over previous approaches is that it only uses a small set of sample images. This method assumes that faces of different individuals have the same shape and different textures. Better rendered results are obtained with this method than when using methods such as the bi-linear model approach [188].

### 6.1.4 Model-Based Approaches

In [189], Principal Component Analysis (PCA) was suggested as a tool for solving the parametric shape-from-shading (SFS) problem. An eigen-head approximation of a 3D head was obtained after training on about 300 laser-scanned range images of real human heads. The ill-posed SFS problem is thereby transformed into a parametric problem, but constant albedo is still assumed. This assumption does not hold for most real face images and it is one of the reasons why most SFS algorithms fail on real face images. To overcome the constant albedo issue, the authors of [68, 182] proposed using a varying albedo reflectance model. They first proposed a new SFS scheme, symmetric SFS. Unlike existing SFS algorithms, Symmetric SFS theoretically allows pointwise 3D information about a symmetric object, represented by the shape gradients $(p, q)$, to be uniquely recovered from a single 2D image. The Symmetric SFS algorithm represents albedo information in the form of a *self-ratio* image, defined as

$$r_I[x,y] = \frac{I_-[x,y]}{I_+[x,y]} = \frac{p[x,y]P_s}{1+q[x,y]Q_s}, \tag{15}$$

where $I$ is the image and is related to the light source $(P_s, Q_s)$ by

$$I[x,y] = \rho[x,y]\frac{1+p[x,y]P_s+q[x,y]Q_s}{\sqrt{1+p[x,y]^2+q[x,y]^2}.\sqrt{1+P_s^2+Q_s^2}} \tag{16}$$

so that recovery of the albedo $\rho$ is not necessary. However, in practical face recognition applications the implementation of this approach is not robust enough. A direct 2D-to-2D approach using a generic 3D model has therefore been proposed. Let the prototype image $I_p$ with $\alpha = 0$ be

$$I_p[x,y] = \rho\frac{1}{\sqrt{1+p^2+q^2}}. \tag{17}$$

Comparing (16) and (17), we obtain

$$I_p[x,y] = \frac{K}{2(1+qQ_s)}(I[x,y]+I[-x,y]), \tag{18}$$

where $K = \sqrt{1+P_s^2+Q_s^2}$. This simple equation directly relates the prototype image $I_p$ to $I[x,y]+I[x,-y]$ which is already available. This direct computation of $I_p$ from $I$ offers the following advantages over the traditional two-step procedure:

- There is no need to recover the varying albedo $\rho[x,y]$.

- There is no need to recover the full shape gradients $(p,q)$.

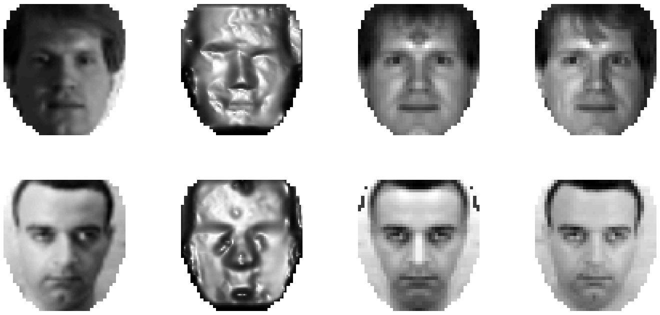


Figure 15: Image rendering comparison. The original images are shown in the first column. The second column shows prototype images rendered using the local SFS algorithm. Prototype images rendered using symmetric SFS are shown in the third column. Finally, the fourth column shows real images that are close to the prototype images.

The rationale behind this method is the observation that all faces have a similar 3D shape; hence the required $q$ can be obtained from a generic model. Using this approach very good prototype images synthesized from front-view input images have been obtained using two publicly available databases, the Yale and Weizmann databases. These databases contain 15 and 24 persons, respectively; each person is represented by four images obtained under different illuminations.

In order to achieve a fully automatic system, light source estimation is needed. After reviewing existing source-from-shading methods, the authors proposed a new model-based symmetric source-from-shading algorithm. Basically it can be formulated as a minimization problem:

$$(\alpha^*, \tau^*) = \arg_{\alpha,\tau} \min(r_{I_{M_F}}(\alpha, \tau)) - r_I)^2. \tag{19}$$

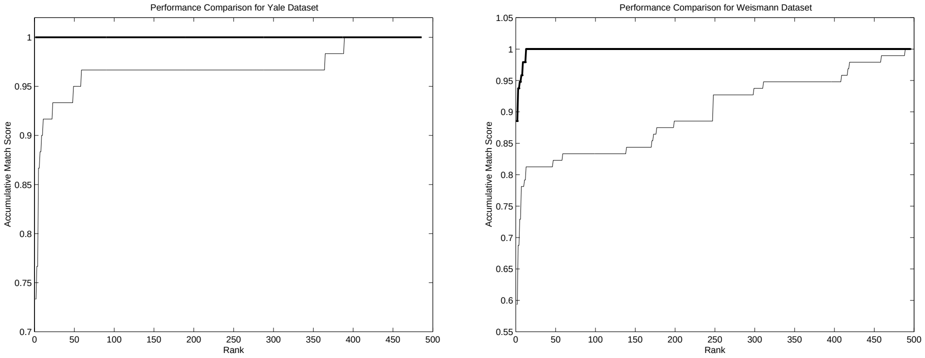
where $r_I$ is the self-ratio image, and $r_{I_{M_F}}$ is the self-ratio image generated from the 3D face model $M_F$ given the hypothesized light source direction represented by $\alpha$ and $\tau$. One advantage of using a 3D face model is that both attached-shadow and cast-shadow effects can be handled.

Figure 15 shows some comparisons between rendered images obtained using this method and using a local SFS algorithm [190]. Significant performance improvements have been reported when the prototype images are used in a subspace LDA system in place of the original input images (Fig. 16).

### 6.2 The pose problem in face recognition

The performance of face recognition systems also drops significantly when pose variations are present in the input images. This difficulty is clearly revealed in the most recent FERET test report, and solving the rotation-in-depth problem has been suggested as a major research issue [3]. When illumination variation is also present the task of face recognition becomes even more difficult (Fig. 17).

Researchers have proposed various methods of handling the rotation problem. They can be divided into three types: 1) Methods in which multiple database images of each



(a) (b)

Figure 16: Enhancing subspace LDA. The thin lines represent the cumulative scores obtained by applying subspace LDA to the original images, while the thick lines represent the scores obtained by applying it to the prototype images. The curves in (a) are for the Yale face database, and those in (b) are for the Weizmann database.

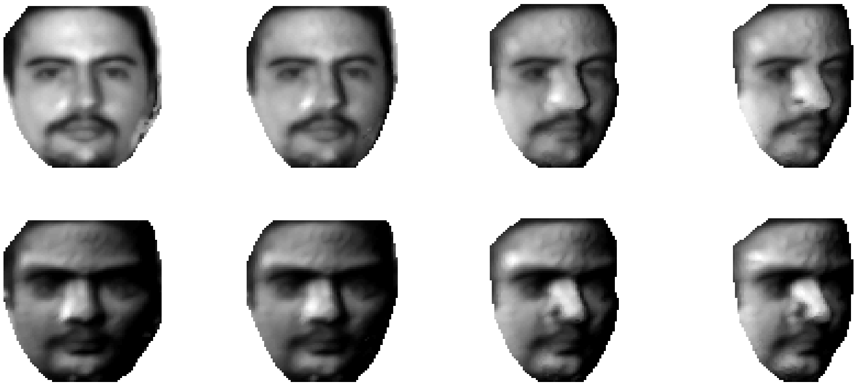


Figure 17: The same face appears different under different poses and illuminations.

person are available [191, 192, 193, 194], 2) hybrid methods when multiple images are available during training but only one database image per person is available during recognition [80, 195, 196, 197], and 3) single image based methods when no training is carried out. The second approach seems to be the most popular one; the third approach does not seem to have received much attention.

### 6.2.1 Multi-Image Based Approaches

One of the earliest examples of the first class of approaches is [193], where a template-based correlation matching scheme is proposed. In this work, pose estimation and face recognition are coupled in an iterative loop. For each hypothesized pose, the input image is aligned to database images corresponding to that pose. The alignment is first carried out via a 2D affine transformation based on three key feature points (eyes and nose), and then optical flow is used to refine the alignment of each template. After this step, the correlation scores of all pairs of matching templates are used for recognition. The main limitations on this method are 1) many different views per person are needed in the database, 2) no lighting variations or facial expressions are allowed, and 3) the computational cost is high since iterative searching is involved. More recently, an illumination-based image synthesis method [194] has been proposed to handle both pose and illumination problems. This method is based on the illumination cone approach [185]. It can handle illumination variation quite well. To handle variations due to rotation, it needs to completely resolve the GBR (generalized-bas-relief) ambiguity when reconstructing the 3D shape.

### 6.2.2 Hybrid Approaches

Numerous algorithms of the second type have been proposed. These methods, which make use of prior class information, are the most successful and practical methods up to now. We review three representative examples here; the first is a linear class based method [196], the second is a graph matching based method [79], and the third is a view-based eigenface approach [45]. The image synthesis method in [196] is based on the assumption of linear 3D object classes and the extension of linearity to images that are 2D projections of the 3D objects. It extends the linear shape model (which is very similar to the active shape model of [198]) from a representation based on feature points to full images of objects. To implement this method, a correspondence between images of the input object and a reference object is established using optical flow. Correspondences between the reference image and other example images having the same pose are also computed. Finally, the correspondence field for the input image is linearly decomposed into the correspondence fields for the examples. Compared to the parallel deformation scheme in [195], this method reduces the need to compute correspondences between images of different poses. This method is extended in [197] to include an additive error term for better synthesis. In [79], a robust face recognition scheme based on EBGM is proposed. The authors assume a planar surface patch at each feature point (landmark), and learn the transformations of "jets" under face rotation. Their results demonstrate substantial improvement in face recognition under rotation. Their method is also fully automatic, including face localization, landmark detection, and flexible graph matching. The drawback of this method is its requirement for accurate landmark localization, which is not an easy task, especially when illumination variations are present. The popular eigenface approach [44] to face recognition has been extended to a view-based eigenface method in order to achieve pose-invariant recognition [45]. This method explicitly codes the pose information by constructing an individual eigenface for each pose. More recently, a unified framework called the bilinear model was proposed in [188]. Despite their popularity, these methods have some common drawbacks: 1) they need many example images to cover all possible views, and 2) the illumination problem is separated from the pose problem.

### 6.2.3 Single Image Based Approaches

Finally, the third class of approaches includes low-level feature based methods, invariant feature based methods, and 3D model based methods. In [51], a Gabor wavelet based feature extraction method is proposed for face recognition which is robust to small-angle rotations. There are many papers on invariant features in the computer vision literature, but to our knowledge, serious application of this technology to face recognition has not yet been explored. However, it is worth pointing out that some recent work on invariant methods based on images [199] may lead to progress in this direction. 3D face models have been used for synthesizing face images under different appearances/lightings/expressions in the computer graphics, computer vision, and model-based coding communities [81, 122, 200]. In these methods, face shape is usually represented by either a polygonal model or a mesh model which simulates tissue. Due to its complexity and computational cost, no serious attempt to apply this technology to face recognition has been made except for [81]. In [201], a unified approach was proposed to solving the pose and illumination problems. This method is a natural extension of the method proposed in [182] to handle the illumination problem. Using this method, input images can be converted into prototype images and then input into existing systems. More specifically, using a generic 3D model, we can approximately solve the correspondence problem required in a 3D rotation and arrive at a direct image-to-image computation:

$$\begin{aligned} I^\theta[x', y'] \;&=\; 1_{z,\theta} I_p[x,y](\cos\theta - p[x,y]\sin\theta)\tfrac{1}{\sqrt{1+P_s^2+Q_s^2}} \\ &\quad \times[\tan(\theta+\theta_0)P_s + \tfrac{q\cos(\theta_0)}{\cos(\theta+\theta_0)}Q_s + 1], \end{aligned} \tag{20}$$

where $1_{z,\theta}$ is the indicator function indicating possible occlusion determined by the shape and rotation angle, the single light source is $(P_s, Q_s, 1)$, and the image (rotated in the $x$-$z$ plane about the $y$-axis) is $I^\theta[x', y']$. As in the pure illumination case, we need to estimate the light source. In addition, pose estimation of the 3D face is needed. To address the varying albedo issue, we again use the self-ratio image and propose the following combined estimation problem (including the pose $\theta$):

$$(\theta^*, \alpha^*, \tau^*) = \arg_{\theta,\alpha,\tau}\min[r_{I_{M_F}}(\alpha,\tau) - r_{I^F}(\theta,\alpha,\tau)]^2, \tag{21}$$

where $r_{I^F(\theta,\alpha,\tau)}$ is the self-ratio image for the virtual frontal view generated from the original image $I_R$ via image warping and texture mapping. For further details, see [201]. Image synthesis examples are shown in Fig. 18.

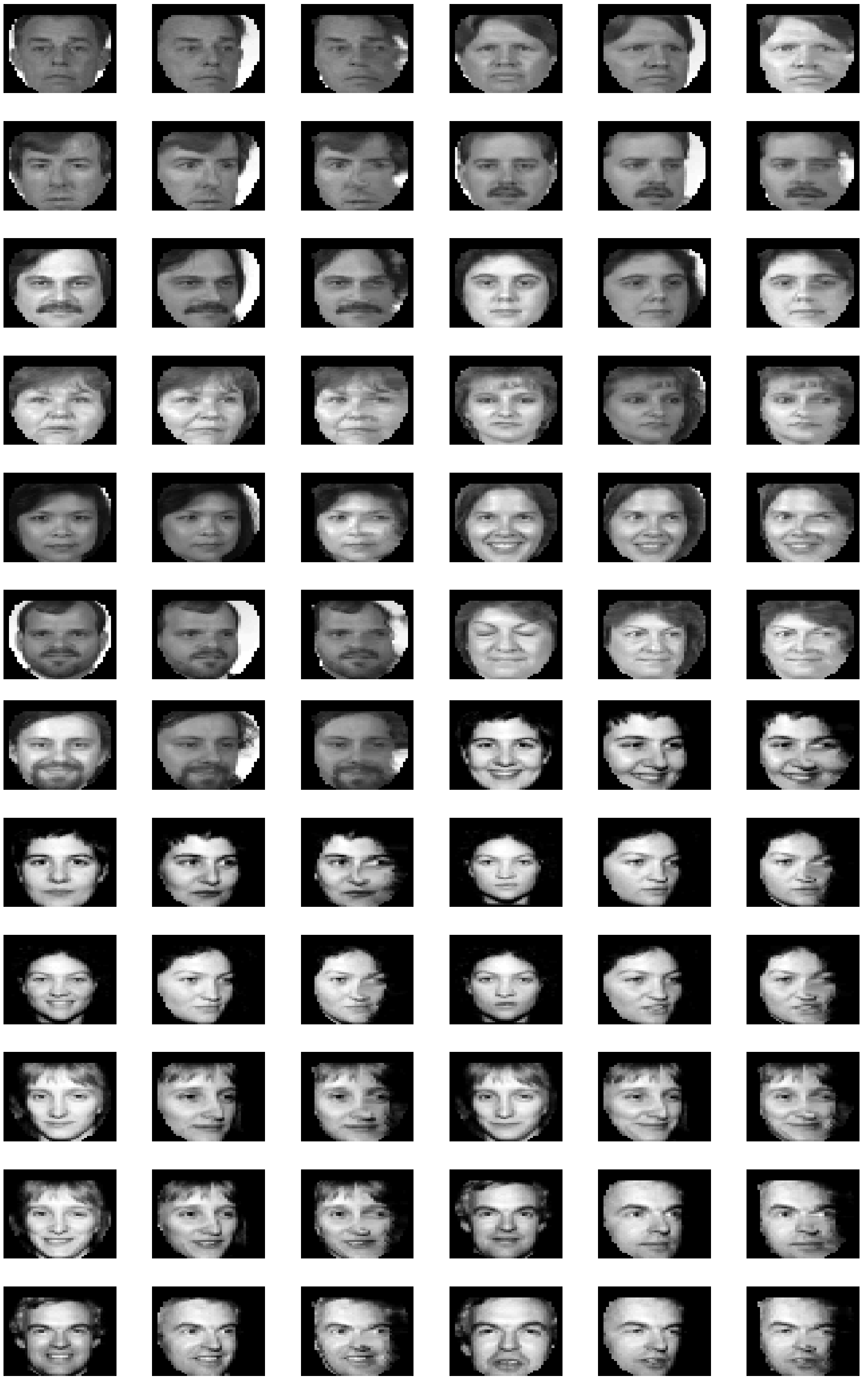


Figure 18: Synthesis of a virtual frontal view from another view: The first column shows the frontal view, the second column shows the rotated view, and the third column shows the virtual frontal view.

## 7 Summary and Conclusions

In this paper, we have presented an extensive survey of machine recognition of human faces. We have focused on segmentation, feature extraction and recognition aspects of the face recognition problem, using information drawn from intensity and range images of faces. In addition, face recognition from image sequences has been reviewed, including basic techniques used in video-based face processing, tracking, modeling, and non-face-based recognition. To emphasize the importance of system evaluation two protocols, the FERET and XM2VTS protocols, have been described in full detail. Finally, we have identified two key problems for any face recognition system: the illumination problem and the pose problem, have categorized proposed methods of solving these two problems, and have discussed the pros and cons of these methods.

We give below a concise summary, followed by conclusions, in the same order in which the topics appear in the paper.

- Machine recognition of faces is emerging as an active research area spanning several disciplines such as image processing, pattern recognition, computer vision, and neural networks. There are numerous commercial applications of FRT such as face verification based ATM and access control, while law enforcement applications include video surveillance. Due to its user-friendly nature, face recognition remains attractive despite the existence of extremely reliable methods of biometric personal identification such as fingerprint analysis and iris scans.

- Over thirty-five years of research in psychophysics and neurosciences on human recognition of faces is documented in the literature. Although we do not feel that machine recognition of faces should strictly follow what is known about human recognition of faces, it is beneficial for engineers who design face recognition systems to be aware of the relevant findings, for example, lighting effects. On the other hand, better machine systems can provide better tools to conduct studies in psychology and neuroscience.

- Segmentation of a face region from a still image or a video is the first key step in a fully automatic face recognition system. During the past five years, significant achievements have been made in this area. Two representative approaches are neural network based systems and example-based learning systems. Face segmentation also has potential application in human-computer interfaces and surveillance systems.

- Both global and local face descriptions are useful. The most significant global descriptions are based on the KL expansion. Local descriptors are derived from regions that contain the eyes, mouth, nose, etc. For better local feature detection, domain knowledge such as face shape should be used.

- Face recognition methods based on sensor modalities such as range images, sketches and infrared images are interesting, but are hard to apply in practice. Many methods have been proposed for face recognition based on image intensities [9]. Basically they can be divided into holistic template matching based systems [33, 44, 56, 58, 59, 60], geometrical local-feature-based schemes [51, 171], and hybrid methods [45]. Even though all these types of systems have been successfully used for to face recognition, they have advantages and disadvantages. Thus appropriate schemes should be chosen based on the specific requirements of a given task. For example, the EBGM-based system [171] has very good performance in general. However, it requires a large-size image, e.g., 128 × 128. This severely restricts its application to video-based surveillance, where the face image size is very small. On the other hand, the subspace LDA system [63] works well with both large and small images, e.g., 96 × 84 or 24 × 21.

- Recognition of faces in a video sequence (especially, in a surveillance video) is still the most challenging problem in face recognition, because the video is of low quality and the images are small. Nevertheless, video-based systems using multiple cues have demonstrated good results in relatively controlled environments

- During the past five years, tracking, modeling and non-face-based recognition of hand gestures and human behavior have been actively studied. One of the reason for this is that generic description of human behavior, not particular to an individual, is both interesting and useful.

- A crucial step in face recognition is the evaluation and benchmarking of numerous algorithms. Two of the most important face databases and their associated evaluation methods are reviewed: the FERET protocol and the XM2VTS protocol. The availability of these protocols has had a significant impact on progress in the development of face recognition algorithms.

- Though many face recognition techniques have been proposed and have demonstrated significant promise, robust face recognition is still difficult. There are at least two major challenges: the illumination and pose problems. An extensive review of methods proposed for solving these problems has been presented and the pros and cons of these methods have also been pointed out. Some difficult issues still remain to be addressed — for example, the problem of aging.

- Methods of multi-modal recognition are needed, though some initial work in this direction has been done. Such methods include the fusion of face recognition with information about speech, iris patterns, fingerprints, and gait.

## References

[1] Proceedings of the International Conferences on Automatic Face and Gesture Recognition, 1995-1998.

[2] Proceedings of the International Conferences on Audio- and Video-Based Person Authentication, 1997,1999.

[3] P.J. Phillips, P. Rauss, and S. Der, "FERET (Face Recognition Technology) Recognition Algorithm Development and Test Report," Technical Report ARL-TR 995, U.S. Army Research Laboratary.

[4] P.J. Phillips, H. Moon, P. Rauss, and S.A. Rizvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 137-143, 1997.

[5] S.A. Rizvi, P.J. Phillips, and H. Moon, "A Verification Protocol and Statistical Performance Analysis for Face Recognition Algorithms," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 833-838, 1998.

[6] P.J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET Testing Protocol," in *Face Recognition: From Theory to Applications* (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie and T.S. Huang, eds.), Berlin: Springer-Verlag, pp. 244-261, 1998.

[7] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 72-77, 1999.

[8] P.J. Phillips, R.M. McCabe, and R. Chellappa, "Biometric Image Processing and Recognition", in *Proceedings, European Signal Processing Conference*, 1998.

[9] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and Machine Recognition of Faces, A Survey," *Proc. IEEE*, Vol. 83, pp. 705-740, 1995.

[10] A. Samal and P. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," *Pattern Recognition*, Vol. 25, pp. 65–77, 1992.

[11] P.K. Kalocsai, W. Zhao, and E. Elagin, "Face Similarity Space as Perceived by Humans and Artifical Systems," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 177-180, 1998.

[12] A. Johnston, H. Hill, and N. Carman, "Recognizing Faces: Effects of Lighting Direction, Inversion and Brightness Reversal," *Cognition*, Vol. 40, pp. 1-19, 1992.

[13] H. D. Ellis, "Introduction to Aspects of Face Processing: Ten Questions in Need of Answers," in *Aspects of Face Processing* (H. Ellis, M. Jeeves, F. Newcombe, and A. Young, eds.), Dordrecht: Nijhoff, pp. 3–13, 1986.

[14] I. Biederman and P. Kalocsai, "Neural and Psychophysical Analysis of Object and Face Recognition," in *Face Recognition: From Theory to Applications* (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie and T.S. Huang, eds.), Berlin: Springer-Verlag, pp. 3-25, 1998.

[15] V. Bruce, P.J.B. Hancock, and A.M. Burton, "Human Face Perception and Identification," in *Face Recognition: From Theory to Applications* (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie and T.S. Huang, eds.), Berlin: Springer-Verlag, pp. 51-72, 1998.

[16] J.C. Bartlett and J. Searcy, "Inversion and Configuration of Faces," *Cognitive Psychology*, Vol. 25, pp.281-316, 1993.

[17] P. Thompson, "Margaret Thatcher — A New Illusion," *Perception*, Vol. 9, pp. 483-484, 1980.

[18] V. Bruce, "Perceiving and Recognizing Faces," *Mind and Language*, pp. 342–364, 1990.

[19] D. Perkins, "A Definition of Caricature and Recognition," *Studies in the Anthropology of Visual Communication*, Vol. 2, pp. 1–24, 1975.

[20] L. D. Harmon, "The Recognition of Faces," *Scientific American*, Vol. 229, No. 5, pp. 71–82, 1973.

[21] A. P. Ginsburg, "Visual Information Processing Based on Spatial Filters Constrained by Biological Data," AMRL Technical Report, pp. 78–129, 1978.

[22] J. Sergent, "Microgenesis of Face Perception," in *Aspects of Face Processing* (H. D. Ellis, M. A. Jeeves, F. Newcombe, and A. Young, eds.), Dordrecht: Nijhoff, 1986.

[23] M.J. Tarr and H.H Bulthoff, "Is Human Object Recognition Better Described by Geon Structural Descriptions or by Multiple Views — Comment on Biederman and Gerhardstein (1993)," *Journal of Experimental Psychology: Human Perception and Performance,* Vol. 21, pp. 71-86, 1995.

[24] I. Biederman, "Recognition by Components: A Theory of Human Image Understanding," *Psychological Review,* Vol. 94, pp. 115-147, 1987.

[25] D. Marr, *Vision*, San Francisco: W. H. Freeman, 1982.

[26] H. Hill, P.G. Schyns, and S. Akamatsu, "Information and Viewpoint Dependence in Face Recognition," *Cognition*, Vol. 62, pp. 201-222, 1997.

[27] H. Hill, V. Bruce, "Effects of Lighting on Matching Facial Surfaces," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, pp. 986-1004, 1996.

[28] B. Knight and A. Johnston, "The Role of Movement in Face Recognition," *Visual Cognition*, Vol. 4, pp. 265-274, 1997.

[29] V. Bruce, *Recognizing Faces*, London: Lawrence Erlbaum Associates, 1988.

[30] K. Sung and T. Poggio, "Example-based Learning for View-based Human Face Detection," A.I. Memo 1521, MIT A.I. Laboratory, 1994.

[31] T.K. Leung, M.C. Burl, and P. Perona, "Finding Faces in Cluttered Scene using Random Labeled Graph Matching," in *Proceedings, International Conference on Computer Vision,* pp. 637-644, 1995.

[32] K.C. Yow and R. Cipolla, "Detection of Human Faces Under Scale, Orientation and Viewpoint Variations," in *Proceedings, International Conference on Automatic Face and Gesture Recognition*, pp. 295-300, 1996.

[33] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 696-710, 1997.

[34] A.J. Colmenarez and T.S. Huang, "Face Detection with Information-Based Maximum Discriminant," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 782-787, 1997.

[35] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian Detection Using Wavelet Templates", in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 193-199, 1997.

[36] N. Kruger, M. Potzsch, and C.v.d. Malsburg, "Determination of Face Position and Pose with a Learned Representation Based on Labelled Graphs," *Image and Vision Computing,* Vol. 15, pp. 665-673, 1997.

[37] E. Osuna, R. Freund, and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 130-136, 1997.

[38] H.A. Rowley, S. Baluja, and T. Kanade, "Neural Network Based Face Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 20, 1998.

[39] H.A. Rowley, S. Baluja, and T. Kanade, "Rotational Invariant Neural Network-Based Face Detection," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 38-44, 1998.

[40] B.K. Low and E. Hjelmas, "Face Detection: A Survey," submitted to *Pattern Recognition*, 1999.

[41] V.N. Vapnik, *The Nature of Statistical Learning Theory,* New York: Springer-Verlag, 1995.

[42] L. Sirovich and M. Kirby, "Low-dimensional Procedure for the Characterization of Human Face," *Journal of the Optical Society of America*, Vol. 4, pp. 519–524, 1987.

[43] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 12, 1990.

[44] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, Vol. 3, pp. 72-86, 1991.

[45] A. Pentland, B. Moghaddam, and T. Starner, "View-based and Modular Eigenspaces for Face Recognition," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[46] R. Bellman, *Introduction to Matrix Analysis*, New York: McGraw-Hill, 1960.

[47] D. Reisfeld and Y. Yeshurun, "Robust Detection of Facial Features by Generalized Symmetry," in *Proceedings, International Conference on Pattern Recognition*, pp. 117–120, 1992.

[48] R. Baron, "Mechanisms of Human Facial Recognition," *International Journal of Man-Machine Studies*, Vol. 15, pp. 137–178, 1981.

[49] P. W. Hallinan, "Recognizing Human Eyes," in *SPIE Proceedings, Vol. 1570: Geometric Methods in Computer Vision*, pp. 214–226, 1991.

[50] H. Moon, R. Chellappa, and A. Rosenfeld, "Optimal Edge-Based Shape Detection," *IEEE International Conference on Image Processing*, Vancouver, BC, Canada, Sept. 2000.

[51] B. S. Manjunath, R. Chellappa, and C. v. d. Malsburg, "A Feature Based Approach to Face Recognition," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 373–378, 1992.

[52] W. Zhao, "Improving the Robustness of Face Recognition," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 78-83, 1999.

[53] K. Jonsson, J. Matas, and J. Kittler, "Learning Salient Features for Real-Time Face Verification," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 60-65, 1999.

[54] J. Matas, K. Messer, and J. Kittler, "Acquisition of a Large Database for Biomeatric Identity Verification," in *BIOSIGNAL 98,* Technical University of Brno, Czech Republic, 1998.

[55] R.A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*,Vol. 7, pp. 179-188.

[56] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, pp. 831-836, 1996.

[57] D.L. Swets and J. Weng, "Discriminant Analysis and Eigenspace Partition Tree for Face and Object Recognition from Views," in *Proceedings, International Conference on Automatic Face and Gesture Recognition*, pp. 192-197, 1996.

[58] P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 711-720, 1997.

[59] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *Journal of the Optical Society of America A*, Vol. 14, pp. 1724-1733, 1997.

[60] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant Analysis of Principal Components for Face Recognition," in *Proceedings, International Conference on Automatic Face and Gesture Recognition*, pp. 336-341, 1998.

[61] W. Zhao, R. Chellappa, and N. Nandhakumar, "Empirical Performance Analysis of Linear Discriminant Classifiers," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 164-169, 1998.

[62] W. Zhao, "Subspace Methods in Object/Face Recognition," in *Proceedings, International Joint Conference on Neural Networks*, 1999.

[63] W. Zhao,, R. Chellappa, and P.J. Phillips, "Subspace Linear Discriminant Analysis for Face Recognition," Technical Report CAR-TR-914, Center for Automation Research, University of Maryland, 1999.

[64] K. Fukunaga, *Statistical Pattern Recognition*, New York: Academic Press, 1989.

[65] S.S. Wilks, *Mathematical Statistics,* New York: Wiley, 1962.

[66] A. Shashua, "Geometry and Photometry in 3D Visual Recognition," PhD Thesis, MIT, 1994.

[67] P. Hallinan, "A Low-Dimensional Representation of Human Faces for Arbitrary Lighting Conditions," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 995-999, 1994.

[68] W. Zhao, *Robust Image Based 3D Face Recognition*, PhD Thesis, University of Maryland, 1999.

[69] "Description of MPEG-7 Content Set", N2467, 1998.

[70] W. Zhao, D. Bhat, N. Nandhakumar, and R. Chellappa, "A Reliable Descriptor for Face Objects in Visual Content," *Journal of Signal Processing: Image Communications,* Special Issue on MPEG-7 Proposals.

[71] "Call for Proposals for MPEG-7 Technology", N2469, 1998.

[72] T. Kohonen, *Self-Organization and Associative Memory*, Berlin: Springer-Verlag, 1988.

[73] Y. S. Abu-Mostafa and D. Psaltis, "Optical Neural Computers," *Scientific American*, Vol. 256, pp. 88–95, 1987.

[74] R. Brunelli and T. Poggio, "HyperBF Networks for Gender Classification," in *Proceedings, DARPA Image Understanding Workshop*, pp. 311–314, 1992.

[75] T. Poggio and F. Girosi, "Networks for Approximation and Learning," *Proc. IEEE*, Vol. 78, pp. 1481–1497, 1990.

[76] J. Buhmann, M. Lades, and C. v. d. Malsburg, "Size and Distortion Invariant Object Recognition by Hierarchiacal Graph Matching," in *Proceedings, International Joint Conference on Neural Networks*, pp. 411–416, 1990.

[77] M. Ladesraj, J. Vorbruggen, J. Buhmann, J.Lange, C. v.d. Malsburg, and R. Wurtz, "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. on Computers*, Vol. 42, pp. 300–311, 1993.

[78] L. Wiskott, J.M. Fellous, N. Kruger, and C.v.d. Malsburg, "Face Recognition and Gender Determination," in *Proceedings, International Workshop on Automatic Face and Gesture Recognition*, pp. 92-97, 1995.

[79] L. Wiskott, J.M. Fellous, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 775-779, 1997.

[80] T. Maurer and C.v.d. Malsburg, "Single-View Based Recognition of Faces Rotated in Depth," in *Proceedings, International Workshop on Automatic Face and Gesture Recognition*, pp. 176-181, 1996.

[81] G. Gordon, "Face Recognition Based on Depth Maps and Surface Curvature," in *SPIE Proceedings, Vol. 1570: Geometric Methods in Computer Vision*, pp. 234–247, 1991.

[82] G. G. Gordon and L. Vincent, "Application of Morphology to Feature Extraction for Face Recognition," in *SPIE Proceedings, Vol. 1658: Nonlinear Image Processing*, 1992.

[83] R.G. Uhl and N.d.V. Lobo, "A Framework for Recognizing a Facial Image from A Police Sketch," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 586-593, 1996.

[84] W. Konen, "Comparing Facial Line Drawings with Gray-Level Images: A Case Study on PHANTOMAS," in *Proceedings, International Conference on Artifical Neural Networks,* pp. 727-734, 1996.

[85] J. Wilder, P.J. Phillips, C.H. Jiang, and S. Wiener, "Comparison of Visible and Infra-Red Imagery for Face Recognition," in *Proceedings, International Conference on Automatic Face and Gesture Recognition*, pp. 182-187, 1996.

[86] S. Ullman, *The Interpretation of Visual Motion*, Cambridge, MA: MIT Press, 1979.

[87] T. S. Huang (ed.), *Image Sequence Analysis,* New York: Springer-Verlag, 1981.

[88] Proceedings of the DARPA Image Understanding Workshops, 1984–1998.

[89] Proceedings of the IEEE Workshops on Visual Motion, 1986, 1989, 1991.

[90] A.J. Howell and H. Buxton, "Towards Unconstrained Face Recognition from Image Sequences," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 224-229, 1996.

[91] S.J. McKenna and S. Gong, "Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 177-183, 1997.

[92] H. Wechsler, V. Kakkad, J. Huang, S. Gutta, and V. Chen, "Automatic Video-Based Person Authentication Using the RBF Network," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 85-92, 1997.

[93] J. Steffens, E. Elagin, and H. Neven, "PersonSpotter — Fast and Robust System for Human Detection, Tracking and Recognition," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 516-521, 1998.

[94] S. Stillman, R. Tanawongsuwa, and I. Essa, "A System For Tracking and Recognizing Multiple People with Multiple Cameras," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 96-101, 1999.

[95] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal Person Recognition Using Unconstrained Audio and Video," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 176-181, 1999.

[96] T. Darrell, B. Moghaddam, and A. Pentland, "Active Face Tracking and Pose Estimation in an Interactive Room," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* 1996.

[97] W.T. Freeman, K. Tanaka, J. Ohta, and K. Kyuma, "Computer Vision for Computer Games," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 100-105, 1996.

[98] J. Healy, J.. Seger, and R.W. Picard, "Quantifying Driver Stress: Developing a System for Collecting and Processing Bio-metric Signals in Natural Situations," Technical Report TR-483, MIT Media Laboratory, 1999.

[99] M.C. Chiang and T.E. Boult, “Local Blur Estimation and Super-resolution,” in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 821-826, 1997.

[100] K. Aizawa, T. Komatsu, and T. Saito, “A Scheme for Accquiring Very High Resolution Images using Multiple Cameras,” in *Proceedings, International Conference on Acoustics, Speech and Signal Processing,* vol. 3, pp. 289-292, 1992.

[101] M. Elad and A. Feuer, “Restoration of a Single Super-resolution Image from Several Blurred, Noisy, and Undersampled Measured Images,” *IEEE Trans. on Image Processing,* vol. 6, pp. 1646-1658, 1997.

[102] M. Berthod, H. Shekarforoush, M. Werman, and J. Zerubia, “Reconstruction of High Resolution 3D Visual Information,” in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 654-657, 1994.

[103] T. Jebara, K. Russel, and A. Pentland, “Mixture of Eigenfeatures for Real-Time Structure from Texture,” Technical Report TR-440, MIT Media Laboratory, 1998.

[104] M. Seibert and A. M. Waxman, “Combining Evidence from Multiple Views of 3-D Objects,” in *SPIE Proceedings, Vol 1611: Sensor Fusion IV: Control Paradigms and Data Structures*, 1991.

[105] A. Shio and J. Sklansky, “Segmentation of People in Motion,” in *Proceedings, IEEE Workshop on Visual Motion*, pp. 325–332, 1991.

[106] E. C. Hildreth, *The Measurement of Visual Motion*, Cambridge, MA: MIT Press, 1984.

[107] W. N. Martin and J. K. Aggarwal, *Motion Understanding, Robot and Human Vision*, Boston: Kluwer Academic Publishers, 1988.

[108] Z. Zhang and O. Faugeras, “3D Dynamic Scene Analysis,” in *Vol. 27, Springer Series in Information Sciences* (T. S. Huang, ed.), New York: Springer-Verlag, 1992.

[109] J. Weng, T. S. Huang, and N. Ahuja, “Motion and Structure from Image Sequences,” in *Vol. 29, Springer Series in Information Sciences*, New York: Springer-Verlag, 1993.

[110] H. H. Nagel, “Analysis Techniques for Image Sequences,” in *Proceedings, International Conference on Pattern Recognition*, pp. 186–211, 1978.

[111] H. H. Nagel, “Image Sequences—Ten (Octal) Years — From Phenomenology Towards a Theoretical Foundation,” in *Proceedings, International Conference on Pattern Recognition*, pp. 1174–1185, 1986.

[112] J. K. Aggarwal and K. Nandhakumar, “On the Computation of Motion from Sequences of Images,” *Proc. IEEE,* Vol. 76, pp. 917–935, 1988.

[113] H. Li, P. Roivainen, and R. Forchheimer, “3-D Motion Estimation in Model-Based Facial Image Coding,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 15, pp. 545–555, 1993.

[114] M. Buck and N. Diehl, “Model-Based Image Sequence Coding,” in *Motion Analysis and Image Sequence Processing* (M. I. Sezan and R. L. Lagendijk, eds.), pp. 285–315, Boston: Kluwer Academic Publishers, 1993.

[115] K. Aizawa et al., “Human Facial Motion Analysis and Synthesis with Application to Model-Based Coding,” in *Motion Analysis and Image Sequence Processing* (M. I. Sezan and R. L. Lagendijk, eds.), pp. 317–348, Boston: Kluwer Academic Publishers, 1993.

[116] J. Strom, T. Jebara, S. Basu, and A. Pentland, “Real Time Tracking and Modeling of Faces: An EKF-based Analysis by Synthesis Approach,” Technical Report TR-506, MIT Media Laboratory, 1999.

[117] D. Terzopoulos, A. Witkin, and M. Kass, “Constraints on Deformable Models: Recovering 3-D Shape and Nonrigid Motion,” *Artificial Intelligence*, Vol. 36, pp. 91–123, 1988.

[118] T. S. Huang, “Modeling, Analysis and Visualization of Non-rigid Object Motion,” in *Proceedings, International Conference on Pattern Recognition*, pp. 361–364, 1990.

[119] A. Pentland and B. Horowitz, “Recovery of Non-rigid Motion and Structure,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, pp. 730–742, 1991.

[120] D. Metaxas and D. Terzopoulos, “Recursive Estimation of Shape and Nonrigid Motion,” in *Proceedings, IEEE Workshop on Visual Motion*, pp. 396–311, 1991.

[121] A. Yuille and P. Hallinan, “Deformable Templates,” in *Active Vision* (A. Blake and A. Yuille, eds.), pp. 21-38, Cambridge, MA: MIT Press, 1992.

[122] D. Terzopoulos and K. Waters, “Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 15, pp. 569-579, 1993.

[123] Y. Yacoob and L. S. Davis, “Computing Spatio-Temporal Representations of Human Faces,” in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[124] R.W. Picard, *Affective Computing,* Cambridge, MA: MIT Press, 1997.

[125] E.S. Vyzas and R.W. Picard, “Affective Pattern Classification,” Technical Report TR-473, MIT Media Laboratory, 1998.

[126] A. Azarbayejani, B. Horowitz, and A. Pentland, “Recursive Estimation of Structure and Motion Using Relative Orientation Constraints,” in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 294-299, 1993.

[127] A. Azarbayejani, T. Starner, B. Horowitz, and A. Pentland, "Visually Controlled Graphics," in *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol. 15, pp. 602-604, 1993.

[128] M. Black and Y. Yacoob, "Tracking and Recognizing Facial Expressions in Image Sequences, Using Local Parametrized Models of Image Motion," Technical Report CS-TR-3401, Center For Automation Research, Unversity of Maryland, 1995.

[129] T. Maurer and C.v.d. Malsburg, "Tracking and Learning Graphs and Pose on Image Sequences of Faces," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 176-181, 1996.

[130] G.D. Hager and P.N. Belhumeur, "Real-Time Tracking of Image Regions with Changes in Geometry and Illumination," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* 1996.

[131] M. Black, D. Fleet, and Y. Yacoob, "A Framework for Modelling Appearance Change in Image Sequences," in *Proceedings, International Conference on Computer Vision*, pp. 660-667, 1998.

[132] P. Ekman, "Facial Expressions of Emotion: An Old Controversy and New Findings," *Philosophical Transactions of the Royal Society of London,* Vol. 335, pp. 63-69, 1992.

[133] A.W. Young and H.D. Ellis, Eds., *Handbook of Research on Face Processing,* New York: Elsevier, 1989.

[134] J.N. Bassili, "Emotion Recognition: The Role of Facial Movement and the Relative Importance of Upper and Lower Areas of the Face," *Journal of Personality and Social Psychology,* Vol. 37, pp. 2049-2059, 1979.

[135] M. Rosenblum, Y. Yacoob, and L. Davis, "Human Emotion Recognition from Motion Using a Radial Basis Function Network Architecture," in *Proceedings, Workshop on Motion of Non-rigid and Articulated Objects,* 1994.

[136] I. Essa and A. Pentland, "A Vision System for Observing and Extracting Facial Action Parameters," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 76-83, 1994.

[137] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Publications in Speech Recognition," *Proc. IEEE,* Vol. 77, pp. 257-285, 1999.

[138] J. Schlenzig, E. Hunter, and R. Jain, "Recursive Identification of Gesture Inputs Using Hidden Markov Model," in *Proceedings, Workshop on Applications of Computer Vision,* pp. 187-194, 1994.

[139] "Learning Visual Behavior for Gesture Analysis," in *Proceedings, International Symposium on Computer Vision,* 1995.

[140] T.E. Starner and A. Pentland, "Visual Recognition of American Sign Language Using Hidden Markov Models," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* 1995.

[141] L.W. Campbell, D.A. Becker, A. Azarbayejani, A.F. Bobick, and A. Pentland, "Invariant Features for 3D Gesture Recognition," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 157-162, 1996.

[142] Y. Cui and J. Weng, "Learning-based Hand Sign Recognition," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* 1995.

[143] J. Triesch and C.v.d. Malsburg, "Robust Classification of Hand Posture against Complex Background," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 170-175, 1996.

[144] V.I. Pavlovic, R. Sharma, and T.S. Huang, "Gestural Interface to a Visual Computing Environment for Molecular Biologists," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 30-35, 1996.

[145] T.S. Huang and V. Pavlovic, "Hand Gesture Modeling, Analysis and Synthesis," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* 1995.

[146] D.M. Gavrila and L.S. Davis, "Towards 3D Model-Based Tracking and Recognition of Human Movement: A Multi-View Approach," in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* 1995.

[147] I. Kakadiaris, D. Metaxas, and R. Bajcsy, "Active Part-Decomposition, Shape and Motion Estimation of Articulated Objects: A Physics-Based Approach," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 980-984, 1994.

[148] A.F. Bobick and J. Davis, "Real-Time Recognition of Activity Using Temporal Templates," in *Proceedings, IEEE Workshop on Applications of Computer Vision,* pp. 1233-1251, 1996.

[149] C. Wren, A. Azerbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," in *IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol 19, pp.780-785, 1997.

[150] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 8-15, 1998.

[151] C. Bregler, "Learning and Recognizing Human Dynamics in Video Sequence," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 568-574, 1997.

[152] I. Haritaoglu, D. Harwood, and L. Davis, “$W^4$ — Who, Where, When, What: A Real-Time System for Detecting and Tracking People,” in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 222-227, 1998.

[153] I. Haritaoglu, D. Harwood, and L. Davis, “$W^4$S: A Real-Time System for Detecting and Tracking People in 2.5D,” in *Proceedings, European Conference on Computer Vision,* 1998.

[154] W.E.L. Grimson, C. Stauffer, R. Romano, and L. Lee, “Using Adaptive Tracking to Classify and Monitor Activities in a Site,“ in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition,* pp. 22-29, 1998.

[155] D. G. Stork and M.E. Hennecke, “Speechreading: An Overview of Image Processing, Feature Extraction, Sensory Integration, and Pattern Recognition Techniques,” in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* 1996.

[156] E.D. Petajan, *Automatic Lipreading to Enchance Speech Recognition,* PhD Thesis, University of Illinois, 1984.

[157] A. Pentland and K. Mase, “Lip reading: Automatic Visual Recognition of Spoken Words,” Technical Report 117, MIT Media Laboratory, 1989.

[158] E.D. Petajan and H.P. Graf, “Robust Face Feature Analysis for Automation Speechreading and Character Animation,” in *Proceedings, International Conference on Automatic Face and Gesture Recognition,* pp. 357-362, 1996.

[159] D. G. Stork and M.E. Hennecke, *Speechreading by Humans and Machines,* Berlin: Springer-Verlag, 1996.

[160] Visionics Coporation, *FaceIt Developer Kit Version 2.0,* http://www.visionics.com/

[161] B. Li and R. Chellappa, “Simultaneous Tracking and Verification via Sequential Posterior Estimation,” submitted to *IEEE Conference on Computer Vision and Pattern Recognition,* 2000.

[162] J. Liu and R. Chen, “Sequential Monte Carlo Methods for Dynamic Systems,” *Journal of the American Statistical Association*, Vol. 93, pp. 1031–1041, 1998.

[163] M. Isard and A. Blake, “Contour Tracking by Stochastic Propagation of Conditional Density,” in *Proceedings, European Conference on Computer Vision*, 1996.

[164] G. T. Candela and R. Chellappa, “Comparative Performance of Classification Methods for Fingerprints,” Technical Report, National Institute of Standards and Technology, 1993.

[165] P. J. Grother and G. T. Candela, “Comparison of Handprinted Digit Classifiers,” Technical Report, National Institute of Standards and Technology, 1993.

[166] S. Rizvi, P. J. Phillips, and H. Moon, "The FERET Verification Testing Protocol for Face Recognition Algorithms," *Image and Vision Computing*, to appear.

[167] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, 2000, to appear.

[168] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss, "The FERET Database and Evaluation Procedure for Face Recognition Algorithms," *Image and Vision Computing*, Vol. 16, pp. 295-306, 1998.

[169] B. Moghaddam, C. Nastar, and A. Pentland, "A Bayesian Similarity Measure for Direct Image Matching," in *Proceedings, International Conference on Pattern Recognition,* 1996.

[170] J. Wilder, "Face Recognition Using Transform Coding of Gray Scale Projection and the Neural Tree Network," in *Artificial Neural Networks with Applications in Speech and Vision* (R.J. Mammone, ed.), New York: Chapman Hall, pp. 520-536, 1994.

[171] K. Okada, J. Steffans, T. Maurer, H. Hong, E. Elagin, H. Neven, and C.v.d. Malsburg, "The Bochum/USC Face Recognition System and How it Fared in the FERET Phase III Test," in *Face Recognition: From Theory to Applications* (H. Wechsler, P.J. Phillips, V. Bruce, F.F. Soulie and T.S. Huang, eds.), Berlin: Springer-Verlag, pp. 186-205, 1998.

[172] H. Moon and P. J. Phillips, "Computational and Performance Aspects of PCA-Based Face Recognition Algorithms," *Perception*, to appear.

[173] H. Moon and P. J. Phillips, "Analysis of PCA-Based Face Recognition Algorithms," in *Empirical Evaluation Techniques in Computer Vision* (K. W. Bowyer and P. J. Phillips, eds.), Los Alamitos, CA: IEEE Computer Society Press, pp. 57-71, 1998.

[174] S. Pigeon and L. Vandendorpe, "The M2VTS Multimodal Face Database (Release 1.00)," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 403-409, 1999.

[175] S. Fischer and B. Duc, "Shape Normalization for Face Recognition," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 21-26, 1997.

[176] C. Kotropoulos, I. Pitas, S. Fischer, and B. Duc, "Face Authentication using Morphological Dynamical Link Architecture," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 169-176, 1997.

[177] E.S. Bigun, J. Bigun, B. Duc, H. Bigun, and S. Fischer, "Expert Conciliation for Multi-Modal Person Authentication Systems by Bayesian Statistics," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 291-300, 1997.

[178] B. Duc, G. Maitre, S. Fischer, and J. Bigun, "Person Authentication by Fusing Face and Speech Recognition," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 311-318, 1997.

[179] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-Labial Speaker Verification," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 319-326, 1997.

[180] S. Pigeon and L. Vandendorpe, "Profile Authentication Using a Chamfer Matching Algorithm," in *Proceedings, International Conference on Audio- and Video-Based Person Authentication*, pp. 185-192, 1999.

[181] Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 721-732, 1997.

[182] W. Zhao and R. Chellappa, "Robust Face Recognition using Symmetric Shape-from-Shading," Technical Report CAR-TR-919, Center for Automation Research, University of Maryland, 1999.

[183] D.W.Jacobs, P.N. Belhumeur, and R. Basri, "Comparing Images under Variable Illumination," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 610-617, 1998.

[184] H. Murase and S. Nayar, "Visual Learning and Recognition of 3D Objects from Appearances," *International Journal of Computer Vision,* Vol. 14, pp. 5-25, 1995.

[185] P.N. Belhumeur and D.J. Kriegman, "What is the Set of Images of an Object Under All Possible Lighting Conditions?" in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 52-58, 1997.

[186] A.S. Georghiades, D.J. Kriegman, and P.N. Belhumeur, "Illumination Cones for Recognition Under Variable Lighting: Faces," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 52-58, 1998.

[187] T. Riklin-Raviv and A. Shashua, "The Quotient Image: Class Based Re-rendering and Recognition with Varying Illuminations," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 566-571, 1999.

[188] W.T. Freeman and J.B. Tenenbaum, "Learing Bilinear Models for Two-Factor Problems in Vision," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 554-560, 1997.

[189] J. Atick, P. Griffin, and N. Redlich, "Statistical Approach to Shape from Shading: Reconstruction of Three-Dimensional Face Surfaces from Single Two-dimensional images," *Neural Computation*, Vol. 8, pp. 1321-1340, 1996.

[190] P.S. Tsai and M. Shah, "A Fast Linear Shape from Shading," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 459-465, 1992,

[191] S. Ullman and R. Basri, "Recognition by Linear Combinations of Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13, pp. 992-1006, 1991.

[192] S. Akamatsu, T. Sasaki, H. Fukamachi, N. Masui, and Y. Suenaga, "An Accurate and Robust Face Identification Scheme," in *Proceedings, International Conference on Pattern Recognition,* pp. 217-220, 1992.

[193] D.J. Beymer, "Face Recognition Under Varying Pose," Technical Report 1461, MIT AI Laboratory, 1993.

[194] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "Illumination-Based Image Synthesis: Creating Novel Images of Human Faces Under Differing Pose and Lighting," in *Proceedings, Workshop on Multi-View Modeling and Analysis of Visual Scenes*, pp. 47-54, 1999.

[195] D.J. Beymer and T. Poggio, "Face Recognition from One Example View,“ in *Proceedings, International Conference on Computer Vision,* pp. 500-507, 1995.

[196] T. Vetter and T. Poggio, "Linear Object Classes and Image Synthesis from a Single Example Image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 733-742, 1997.

[197] E. Sali and S. Ullman, "Recognizing Novel 3-D Objects Under New Illumination and Viewing Position Using a Small Number of Example Views or Even a Single View," in *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pp. 153-161, 1998.

[198] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models—Their Training and Application," *Computer Vision and Image Understanding,* Vol. 61, pp. 18-23, 1995.

[199] R. Alferez and Y.F Wang, "Geometric and Illumination Invariants for Object Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, pp. 505-536, 1999.

[200] T. Akimoto, Y. Suenaga, and R.S. Wallace, "Automatic Creation of 3D Facial Models," *IEEE Computer Graphics and Applications*, Vol. 13, pp. 16-22, 1993.

[201] W. Zhao and R. Chellappa, "SFS Based View Synthesis for Robust Face Recognition," in *Proceedings, International Conference on Automatic Face and Gesture Recognition*, 2000.