Probabilistic Visual Learning for Detection and Recognition

by

Baback Moghaddam

B.S., George Mason University (1989)
M.S., George Mason University (1992)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 1997

Author ...
Department of Electrical Engineering and Computer Science
August 8, 1997

Certified by .....
Alex P. Pentland
Toshiba Professor, Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Theses

Probabilistic Visual Learning for Detection and Recognition

by

Baback Moghaddam

Submitted to the Department of Electrical Engineering and Computer Science
on August 8, 1997, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, we present an unsupervised technique for visual learning which is based on density estimation in high-dimensional spaces using an eigenspace decomposition. Two types of density estimates are derived for modeling the training data: a multivariate Gaussian (for unimodal distributions) and a Mixture-of-Gaussians model (for multimodal distributions). These probability densities are then used to formulate a maximum-likelihood estimation framework for automatic target detection as well as a novel Bayesian similarity measure for image matching for image databases. This learning technique has been specifically applied to the problems of detection and recognition of human faces. The resulting automatic face recognition system has been extensively tested by the US Army Research Laboratory as part of ARPA's "FERET" Face Recognition Program, where it was most recently found to be the top competitor.

Thesis Supervisor: Alex P. Pentland
Title: Toshiba Professor, Media Arts and Sciences

Acknowledgments

First of all I would like to thank my thesis advisor Alex Pentland for his help, encouragement and patience during the last 5 years and also my committee members, Prof. Eric Grimson and Prof. Jacob White for their time and commitment.

I would like to thank all my friends in Vismod for their help and support throughout the years. Special thanks go to the TRS-80 Gang: Lee Campbell, Ali Azarbayejani, Christopher Wren, Andy Wilson, Dave Becker, Matt Krom, Claudio Pinhanez and Stephen Intille.

I would especially like to thank my sister Marjan, for her constant encouragement in my moments of despair.

Also great thanks to Wasiuddin Wahid for his help with the FERET tests as well as kind help in being a typist for me.

Finally, I'm greatly indebted to my friend and collaborator Chahab Nastar who contributed his knowledge, expertise and code in the XYI warping method.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Problem

The central problem tackled in this thesis is that of automatic detection and recognition of objects represented by 2D image patterns (mainly frontal human faces) using a probabilistic framework. Given an input image containing a human face, for example, we would like to designate the most likely location of the face in the image, compensate for sources of image variation (such as scale, translation, rotation and lighting) and ultimately recognize the identity of the individual in the image. This "face-finding/recognition" task constitutes a visual routine requiring a detection/segmentation mechanism for visual attention followed by an identification mechanism for visual recognition.

Visual attention is the process of restricting higher-level processing to a subset of the visual field, referred to as the *focus-of-attention* (FOA). Palmer [31] has suggested that visual attention is the process of locating the object of interest and placing it in a *canonical* (or object-centered) reference frame suitable for recognition (or template matching). We have developed a computational technique for automatic object recognition, which is in accordance with Palmer's model of visual attention (see section 4.2.1). The system uses a probabilistic formulation for the estimation of the position and scale of the object in the visual field and remaps the FOA to an object-centered reference frame, which is subsequently used for recognition and

(a) (b) (c) (d)

Figure 1-1: (a) input image, (b) face detection, (c) input image, (d) hand detection

verification.

At a simple level the underlying mechanism of attention during a visual search task can be based on a spatiotopic *detection* map $S(i,j)$ which is a function of the image information in a local region R

$$S(i,j) = f\left[\{I(i+r, j+c) : (r,c) \in R\}\right] \tag{1.1}$$

For example detection maps have been constructed which employ spatio-temporal changes as cues for foveation [1] or other low-level image features such as local symmetry for detection of interest points [40]. However bottom-up techniques based on low-level features lack *context* with respect to high-level visual tasks such as object recognition. In a recognition task, the selection of the FOA is driven by higher-level goals and therefore requires internal representations of an object's appearance and a means of comparing candidate objects in the FOA to the stored object models.

Specifically, in an object-based visual search the detection map is a function of the degree of match between a candidate object in a local image region and an internal model of the object. In view-based recognition (as opposed to 3D geometric or invariant-based recognition), the detection can be formulated in terms of visual similarity using a variety of metrics ranging from simple template matching scores to more sophisticated measures using, for example, robust statistics for image correlation [6].

1.2 The Approach

In this thesis, however, we are primarily interested in detection maps which have a *probabilistic* interpretation as object-class membership functions or *likelihoods*. These likelihood functions are learned by applying density estimation techniques in complementary subspaces obtained by an eigenvector decomposition. Our approach to this learning problem is *view-based* — *i.e.,* the learning and modeling of the visual appearance of the object from a (suitably normalized and preprocessed) set of training imagery. Figure 1-1 shows examples of the automatic selection of FOA for detection of faces and hands. In each case, the target object's probability distribution was *learned* from training views and then subsequently used in computing likelihoods for detection. The face representation is based on appearance (normalized grayscale image) whereas the hand's representation is based on the shape of its contour. The maximum likelihood (ML) estimates of position and scale are shown in the figure by the cross-hairs and bounding box, respectively.

Once the object has been detected, it is normalized for scale, translation and in-plane rotation, prior to visual recognition. The face recognition system developed in this thesis makes use of two types of recognition strategies: the first consists of projecting the normalized image into a face-space eigenspace and performing identity matching using nearest neighbor rule. Whereas the second approach is a more sophisticated method which uses a probabilistic similarity measure based on two types of learned image deformations: intra-personal variations corresponding to different appearances of the same individual and extra-personal variations corresponding to appearance changes between different individuals. Both these classes are represented by their respective probability density functions which are derived from training data using the same eigenvector decomposition method used for visual modeling for object detection.

1.3 The Results

The experimental results in this thesis test both the detection and recogniton aspects of our system. For the detection stage, the visual density estimates are used to formulate a *maximum likelihood* (ML) framework for pattern detection which is used as part of a detection and normalization system for automatic face recognition. This ML detector is then compared in experiments over a standard matched filter type detection scheme where it is shown to perform significantly better (by typically an order of magnitude) for detection of facial features. Additionally, the face processing system has been tested as part of the US Army's FERET face recognition competition where a 97% reliability in detection was obtained.

In addition, we have tested the recognition aspect of our system with various face databases. The nearest neighbor eigenspace matching technique has been tested on the media lab database of 8000 faces where it was found to have a 95% recognition accuracy. This simple matching rule was also tested on the FERET database where it achieved a slightly lower recognition rate (87%). The second recognition strategy, using a probabilistic similarity measure, however was found to give the best overall performance. In September 1996, our system was found to be the top performer in the FERET competition with recognition rates of 95% on a database of size 3000.

1.4 Thesis Outline

The organization of this thesis is as follows. In Chapter 2, we provide background on related work on visual detection and recognition. In Chapter 3, we present a subspace learning method for characterizing the density of high-dimensional visual data. These density estimates are then used in visual detection examples presented in Chapter 4 which includes examples of head detection, facial feature detection as well as hands. This chapter also includes recognition examples using a nearest neighbor matching technique. In Chapter 5, we propose a probabilistic alternative to the nearest neighbor matching and also use a novel representation for image differences

based on a physically deformable XYI surface. Next in Chapter 6, we present results of the FERET face recognition tests in which our face recognition system participated. Finally in Chapter 7, we conclude with a summary of the thesis and discuss future directions for research.

Chapter 2

Background

In recent years, computer vision research has witnessed a growing interest in eigenvector analysis and subspace decomposition methods. In particular, eigenvector decomposition has been shown to be an effective tool for solving problems which use high-dimensional representations of phenomena which are intrinsically low-dimensional. This general analysis framework lends itself to several closely related formulations in object modeling and recognition which employ the *principal modes* or characteristic *degrees-of-freedom* for description. The identification and parametric representation of a system in terms of these principal modes is at the core of recent advances in physically-based modeling [34], correspondence and matching [42], and parametric descriptions of shape [8].

Eigenvector-based methods also form the basis for data analysis techniques in pattern recognition and statistics where they are used to extract low-dimensional subspaces comprised of statistically uncorrelated variables which tend to simplify tasks such as classification. The Karhunen-Loeve Transform (KLT) [23] and Principal Components Analysis (PCA) [16] are examples of eigenvector-based techniques which are commonly used for dimensionality reduction and feature extraction in pattern recognition.

In computer vision, eigenvector analysis of *imagery* has been used for characterization of human faces [20] and automatic face recognition using "eigenfaces" [44][32]. More recently, principal component analysis of imagery has also been applied for

robust object detection [32][7], nonlinear image interpolation [5], visual learning for object recognition [25][47], as well as visual servoing for robotics [30].

Specifically, Murase & Nayar [25] used a low-dimensional *parametric* eigenspace for recovering object identity and pose by matching views to a spline-based hypersurface. Nayar *et al.* [30] have extended this technique to visual feedback control and servoing for a robotic arm in "peg-in-the-hole" insertion tasks. Pentland *et al.* [32] proposed a view-based multiple-eigenspace technique for face recognition under varying pose as well as for the detection and description of facial features. Similarly, Burl *et al.* [7] used Bayesian classification for object detection using a feature vector derived from principal component images. Weng [47] has proposed a visual learning framework based on the KLT in conjunction with an optimal linear discriminant transform for learning and recognition of objects from 2D views.

However, these authors (with the exception of [32]) have used eigenvector analysis primarily as a dimensionality reduction technique for subsequent modeling, interpolation, or classification. In contrast, our method uses an eigenspace decomposition as an integral part of an efficient technique for probability density estimation of high-dimensional data.

Our learning method estimates the *complete* probability distribution of the object using an eigenvector decomposition of the sample covariance matrix of a set of training views. The desired object density is hence decomposed into two components: the density in the principal subspace (containing the traditionally-defined principal components) and its orthogonal complement (which is usually discarded in PCA). We formulate an optimal density estimate for the case of Gaussian data and a near-optimal estimator for arbitrarily complex distributions in terms of a Mixture-of-Gaussians. These density estimates are then used for maximum likelihood detection of faces and articulated hands in natural images.

Furthermore, our learning method differs from *supervised* visual learning with function approximation networks [38] in which a hypersurface representation of an input/output map is automatically learned from a set of examples. Instead, we use a probabilistic formulation which combines the two standard paradigms of *unsupervised*

learning — PCA and density estimation — to arrive at a computationally feasible estimate of the class conditional density function $P(\mathbf{x}|\Omega)$ for an object based on its (high-dimensional) visual appearance — its image $\mathbf{x}$.

2.1 Visual Object Detection

The standard detection paradigm in image processing is that of normalized correlation or template matching. However this approach is only optimal in the simplistic case of a *deterministic* signal embedded in additive white Gaussian noise. When we begin to consider a object *class* detection problem — *e.g,* finding a generic human face or a human hand in a scene — we must incorporate the underlying probability distribution of the object. Subspace methods and eigenspace decompositions are particularly well-suited to such a task since they provide a compact and *parametric* description of the object's appearance and also automatically identify the *degrees-of-freedom* of the underlying statistical variability.

In particular, the eigenspace formulation leads to a powerful alternative to standard detection techniques such as template matching or normalized correlation. The reconstruction error (or residual) of the eigenspace decomposition (referred to as the "distance-from-face-space" in the context of the work with "eigenfaces" [44]) is an effective indicator of similarity [44, 32]. The residual error is easily computed using the projection coefficients and the original signal energy. This detection strategy is equivalent to matching with a linear combination of *eigentemplates* and allows for a greater range of distortions in the input signal (including lighting, and moderate rotation and scale). In a statistical signal detection framework, the use of eigentemplates has been shown to yield superior performance in comparison with standard matched filtering [21][32].

In [32] we used this formulation for a modular eigenspace representation of facial features where the corresponding residual — referred to as "distance-from-*feature*-space" or DFFS — was used for localization and detection. Given an input image, a saliency map was constructed by computing the DFFS at each pixel. When using M

eigenvectors, this requires M convolutions (which can be efficiently computed using an FFT) plus an additional local energy computation. The global minimum of this distance map was then selected as the best estimate of the location of the object.

In this thesis we will show that the DFFS can be interpreted as an estimate of a marginal component of the probability density of the object and that a complete estimate must also incorporate a second marginal density based on a complementary "distance-*in*-feature-space" (DIFS). Using our estimates of the object densities, we formulate the problem of object detection from the point of view of a maximum likelihood (ML) estimation problem. Specifically, given the visual field, we estimate the position (and scale) of the image region which is most representative of the object of interest. Computationally this is achieved by sliding an m-by-n observation window throughout the image and at each location computing the *likelihood* that the local subimage $\mathbf{x}$ is an instance of the object class Ω — *i.e.*, $P(\mathbf{x}|\Omega)$. After this probability map is computed, we select the location corresponding to the highest likelihood as our ML estimate of the object location. Note that the likelihood map can be evaluated over the entire parameter space affecting the object's appearance which can include transformations such as scale and rotation.

2.2 Visual Object Recognition

Current work in the area of image-based object modeling and visual recognition treats the shape and texture components of an object in a separate and often independent manner. The technique of extracting shape and forming a shape-normalized or "shape-free" grayscale component was suggested by Craw & Cameron [10], which used an eigenface technique on shape-free faces for matching and recognition. Recently Craw *et al.* [11] have done a study which combines these two independently derived components (a manually-extracted shape component plus a shape-free texture) for enhanced recognition performance. Similarly, Lanitis *et al.* [22] have developed an automatic face-processing system which is capable of combining the shape and texture components for recognition, albeit independently. Their system detects

canonical points on the face and uses these landmarks to warp faces to a shape-free representation prior to implementing an eigenface technique for characterizing grayscale variations (face texture).

Similarly, the face vectorizer system of Beymer & Poggio [3] uses optical flow to obtain a shape representation decoupled from that of texture (in the form of a 2D correspondence field between a given face and a canonical model). However, one of the difficulties with using optical flow for correspondance between two different individuals is that the technique is inherently failure-prone when there are large grayscale variations between the images (*e.g.,* presence/absence of facial hair). A pixel correspondence technique must be able to deal with intensity variations as well as spatial deformations, preferably in a unified framework.

In this thesis, we use a novel image representation which combines both the spatial (XY) and grayscale (I) components of the image into a 3D surface (or manifold) and then efficiently solves for a dense correspondence field in the XYI space. These image manifolds are modeled as physically-based deformable surfaces which undergo deformations in accordance with a specified force field. The physical dynamics of the system are efficiently solved for using a formulation in terms of the *analytic* modes of vibration [26]. This manifold matching technique can be viewed as a more general formulation for image correspondence which, unlike optical flow, does *not* require a constant brightness assumption. In fact, by simply disabling the I component of our deformations we can obtain a standard 2D deformable mesh which yields correspondences similar to an optical flow technique with thin-plate regularizers.

This novel image correspondence method is used to match two facial images by deforming the XYI surface of one image into the another (under "physical forces" exerted by nearby mesh nodes). The resulting vector of displacements yields a pixel-dense set of correspondences which can be used for image warping. In addition the vector of modal amplitudes is then used to classify the deformation into one of two categories: *interpersonal* vs. *extrapersonal.* This final classification is performed using the *a posteriori* probabilities computed from the two class-conditional likelihoods which are themselves estimated from training data using an efficient subspace method

for density estimation of high-dimensional Gaussian data.

2.3 Related Work

In this section we review some of the related work on face processing and extensions of the eigenspace visual learning techniques to other domains.

2.3.1 Recognition: Cootes, Taylor & Lanitis

Lanitis *et al.* [22] have developed a system which uses an iterative gradient descent process with an Active Shape Model which consists of an eigenspace representation of the XY coordinates of a set of fiducial 2-D points plus local grayscale information. Although impressive face fitting performances have been demonstrated with this technique, it is not clear whether this system can function as a general face spotter. Other applications that have been demonstrated using this system are recovering pose, facial identification, gender recognition, as well as expression recognition.

2.3.2 Matching: Jones & Poggio

A similar approach to the XYI technique presented in this thesis is work by Jones and Poggio [17] which uses a linear combination of shape and texture components of an image for matching. Shape is represented by a linear combination of optical flow warp fields with respect to a reference image plus a global affine transformation. The texture component is represented by a linear combination of normalized textures. Matching is obtained by minimizing the L_2 norm of a novel image with that of the model using a stochastic gradient descent method.

The similarity to the XYI technique is that the shape and texture components are solved for simultaneously. The difference however is that these components are represented independently as opposed to the unified method of XYI warping. Using a hierarchical pyramid technique, Jones demonstrates robustness with respect to scale as well rotation and translation. In the XYI technique, these invariances are obtained

by the coarse alignment provided by the face processor which uses head location and eye locations to align facial images prior to XYI warping.

One of the possible disadvantages of Jones' method is its reliance on optical flow which is unreliable when matching images which have large grayscale variations due to facial hair and/or lighting, necessitating the use of manual correspondences as an initial step.

It would be interesting to see the recognition performance of Jones' method on a standardized test set such as the FERET database since no examples of recognition performance are provided in the cited paper.

2.3.3 Synthesis: Beymer & Poggio

In Beymer's work [3], images are vectorized by computing a correspondence flow field and texture map which are then mapped to a 2-D pose/expression space using a RBF network. The reverse of this analysis network yields a synthesis network which given pose and expression parameters can synthesize a novel view using the established correspondences. An interesting application of Beymer's work is "directing" the pose/expression of another person using the analysis/synthesis network. This technique can also be used to generate synthetic views from one model view with applications towards pose invariant face recognition.

2.3.4 Detection: Sung & Poggio

The system developed by Sung and Poggio [43] uses a distribution-based face model for face spotting. This system uses face and non-face patterns, representing each as a mixture of Gaussians. Two sets of distances are computed to each component for each class and fed to a neural network which is trained to output 0 or 1 depending on the input pattern being a face. Using an extensive training set of various face patterns, which includes small translations, rotations and scale variations and a bootstrapped set of face-like non-face patterns, they are able to achieve an impressive detection performance.

This system was further enhanced by Rowley *et al.* [41] which introduced some further preprocessing consisting of lighting normalization and histogram equalization and which replaced the distribution-based model of Sung with an arbitrated ensemble of networks which were trained to output 0/1 values based directly on the preprocessed image. This results in significant reduction in computational costs over the system of Sung and Poggio.

2.3.5 Extensions

An interesting new extension of eigenspace methods for detection and tracking is the work of Black and Jepson [4] which incorporates robust norms for computing expansion coefficients, a "subspace constancy assumption" which uses parameterized optical flow estimation to obtain the view as well as the affine transformation between the eigenspace and the image. This technique is used to track objects which simultaneously undergo changes of view as well as affine image motions.

Another recent extension of eigenspace methods is the parametric feature detection technique of Nayar *et al.* [29] where various types of low-level features such as step-edge, roof-edge, corners and circular disks are modeled in a parametric eigenspace formulation. These features are detected using the "distance to manifold" metric. The results obtained with this technique show significant improvement over standard edge detection techniques.

Chapter 3

Density Estimation in Eigenspace

Our approach to automatic visual learning is based on density estimation. However, instead of applying estimation techniques directly to the original high-dimensional space of the imagery, we use an eigenspace decomposition to yield a computationally feasible estimate. Specifically, the eigenspace analysis is applied to a set of training views of the object in order to identify a principal subspace which captures the *intrinsic* dimensionality of the data. The component of the complete density in this lower-dimensional subspace is then estimated using a suitable parametric form. In addition, we implicitly model the component of the distribution in the *orthogonal* subspace. The complete density estimate can be efficiently computed from the lower-dimensional principal components. Our density estimate is shown to be *optimal* in the case of Gaussian-distributed training data. We also formulate an approximate (near-optimal) density estimate for more realistic data with arbitrary and multimodal distributions.

Specifically, given a set of training images $\{\mathbf{x}^t\}_{t=1}^{N_T}$, from an object class Ω, we wish to estimate the class membership or *likelihood* function for this data — *i.e.*, $P(\mathbf{x}|\Omega)$. In this section, we examine two density estimation techniques for visual learning of high-dimensional data. The first method is based on the assumption of a Gaussian distribution while the second method generalizes to arbitrarily complex distributions using a Mixture-of-Gaussians density model. Before introducing these estimators we briefly review eigenvector decomposition as commonly used in PCA.

3.1 Principal Component Imagery

Given a training set of m-by-n images $\{I^t\}_{t=1}^{N_T}$, we can form a training set of vectors $\{\mathbf{x}^t\}$, where $\mathbf{x} \in \mathcal{R}^{N=mn}$, by lexicographic ordering of the pixel elements of each image I^t. The basis functions for the KLT [23] are obtained by solving the eigenvalue problem

$$\Lambda = \Phi^T \Sigma \Phi \tag{3.1}$$

where Σ is the covariance matrix given by

$$\Sigma = \frac{1}{N_T} \sum_{i=1}^{N_T} (\mathbf{x}^t - \bar{\mathbf{x}})(\mathbf{x}^t - \bar{\mathbf{x}})^T \tag{3.2}$$

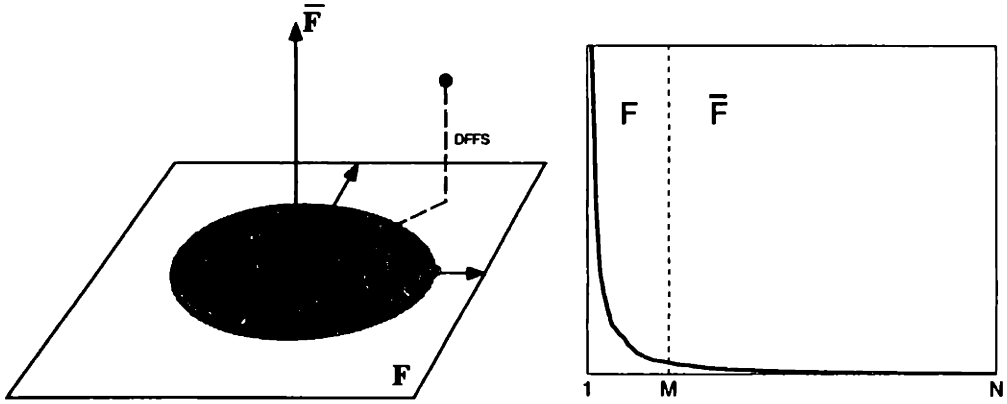
with the mean vector

$$\bar{\mathbf{x}} = \frac{1}{N_T} \sum_{i=1}^{N_T} \mathbf{x}^t \tag{3.3}$$

Thus Φ is the eigenvector matrix of Σ and Λ is the corresponding diagonal matrix of eigenvalues. The unitary matrix Φ defines a coordinate transform (rotation) which *decorrelates* the data and makes explicit the *invariant subspaces* of the matrix operator Σ. In PCA, a partial KLT is performed to identify the largest-eigenvalue eigenvectors and obtain a principal component feature vector $\mathbf{y} = \Phi_M^T \tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$ is the mean-normalized image vector and Φ_M is the submatrix containing the columns of Φ corresponding to the principal eigenvectors. PCA can be seen as a linear transformation $\mathbf{y} = \mathcal{T}(\mathbf{x}) : \mathcal{R}^N \rightarrow \mathcal{R}^M$ which extracts a lower-dimensional subspace of the KL basis corresponding to the maximal eigenvalues. These principal components preserve the major linear correlations in the data and discard the minor ones.[1]

By ranking the eigenvectors of the KL expansion with respect to their eigenvalues and selecting the first M principal components we form an orthogonal decomposition of the vector space $\mathcal{R}^N$ into two mutually exclusive and complementary subspaces: the principal subspace (or feature space) $F = \{\Phi_i\}_{i=1}^M$ containing the principal component

[1] In practice the number of training images N_T is far less than the dimensionality of the imagery N, consequently the covariance matrix Σ is singular. However, the first $M < N_T$ eigenvectors can always be computed (estimated) from N_t samples using, for example, a Singular Value Decomposition [13].



(a) (b)

Figure 3-1: (a) Decomposition into the principal subspace F and its orthogonal complement $\bar{F}$ for a Gaussian density, (b) a typical eigenvalue spectrum obtained from PCA.

(first M columns of Φ) and its orthogonal complement $\bar{F} = \{\Phi_i\}_{i=M+1}^{N}$ (the remaining columns). This orthogonal decomposition is illustrated in Figure 3-1(a) where we have a prototypical example of a distribution which is embedded entirely in F. In practice there is always a signal component in $\bar{F}$ due to the minor statistical variabilities in the data or simply due to the observation noise which affects every element of $\mathbf{x}$.

In a partial KL expansion, the residual reconstruction error is defined as

$$\epsilon^2(\mathbf{x}) = \sum_{i=M+1}^{N} y_i^2 = ||\tilde{\mathbf{x}}||^2 - \sum_{i=1}^{M} y_i^2 \tag{3.4}$$

and can be easily computed from the first M principal components and the L_2 norm of the mean-normalized image $\tilde{\mathbf{x}}$. Consequently the L_2 norm of every element $\mathbf{x} \in \mathcal{R}^N$ can be decomposed in terms of its projections in these two subspaces. We refer to the component in the orthogonal subspace $\bar{F}$ as the "distance-from-feature-space" (DFFS) which is a simple Euclidean distance and is equivalent to the residual error $\epsilon^2(\mathbf{x})$ in Eq.(3.4). The component of $\mathbf{x}$ which lies *in* the feature space F is referred to as the "distance-in-feature-space" (DIFS) but is generally not a distance-based norm, but can be interpreted in terms of the probability distribution of y in F.

3.2 Gaussian Densities

We begin by considering an optimal approach for estimating high-dimensional Gaussian densities. We assume that we have (robustly) estimated the mean $\bar{\mathbf{x}}$ and covariance Σ of the distribution from the given training set $\{\mathbf{x}^t\}$.[2] Under this assumption, the likelihood of an input pattern $\mathbf{x}$ is given by

$$P(\mathbf{x}|\Omega) = \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^T\Sigma^{-1}(\mathbf{x}-\bar{\mathbf{x}})\right]}{(2\pi)^{N/2}\,|\Sigma|^{1/2}} \tag{3.5}$$

The sufficient statistic for characterizing this likelihood is the *Mahalanobis* distance

$$d(\mathbf{x}) \;=\; \tilde{\mathbf{x}}^T\Sigma^{-1}\tilde{\mathbf{x}} \tag{3.6}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$. However, instead of evaluating this quadratic product explicitly, a much more efficient and robust computation can be performed, especially with regard to the matrix inverse Σ^{-1}. Using the eigenvectors and eigenvalues of Σ we can rewrite Σ^{-1} in the diagonalized form

$$\begin{aligned} d(\mathbf{x}) &= \tilde{\mathbf{x}}^T\Sigma^{-1}\tilde{\mathbf{x}} \\ &= \tilde{\mathbf{x}}^T\left[\Phi\Lambda^{-1}\Phi^T\right]\tilde{\mathbf{x}} \\ &= \mathbf{y}^T\Lambda^{-1}\mathbf{y} \end{aligned} \tag{3.7}$$

where $\mathbf{y} = \Phi^T\tilde{\mathbf{x}}$ are the new variables obtained by the change of coordinates in a KLT. Because of the diagonalized form, the *Mahalanobis* distance can also be expressed in terms of the sum

$$d(\mathbf{x}) \;=\; \sum_{i=1}^{N}\frac{y_i^2}{\lambda_i} \tag{3.8}$$

In the KLT basis, the *Mahalanobis* distance in Eq.(3.6) is conveniently *decoupled* into a weighted sum of uncorrelated component energies. Furthermore, the likelihood becomes a *product* of independent separable Gaussian densities. Despite its

[2] In practice, a full rank N-dimensional covariance Σ can not be estimated from N_T independent observations where typically $N_T << N$. But as we shall see our estimator does not require the full covariance, but only its first M principal eigenvectors where $M < N_T << N$.

simpler form, evaluation of Eq.(3.8) is still computationally infeasible due to the high-dimensionality. We therefore seek to *estimate* $d(\mathbf{x})$ using only M projections. Intuitively, an obvious choice for a lower-dimensional representation is the principal subspace indicated by PCA which captures the major degrees of statistical variability in the data.[3] Therefore, we divide the summation into two independent parts corresponding to the principal subspace $F = \{\Phi_i\}_{i=1}^{M}$ and its orthogonal complement $\bar{F} = \{\Phi_i\}_{i=M+1}^{N}$

$$d(\mathbf{x}) = \sum_{i=1}^{M} \frac{y_i^2}{\lambda_i} + \sum_{i=M+1}^{N} \frac{y_i^2}{\lambda_i} \tag{3.9}$$

We note that the terms in the first summation can be computed by projecting $\mathbf{x}$ onto the M-dimensional principal subspace F. The remaining terms in the second sum $\{y_i\}_{i=M+1}^{N}$, however, can not be computed explicitly in practice because of the high-dimensionality. However, the *sum* of these terms is available and is in fact the DFFS quantity $\epsilon^2(\mathbf{x})$ which can be computed from Eq.(3.4). Therefore, based on the available terms, we can formulate an estimator for $d(\mathbf{x})$ as follows

$$\begin{aligned} \hat{d}(\mathbf{x}) &= \sum_{i=1}^{M} \frac{y_i^2}{\lambda_i} + \frac{1}{\rho}\left[\sum_{i=M+1}^{N} y_i^2\right] \\ &= \sum_{i=1}^{M} \frac{y_i^2}{\lambda_i} + \frac{\epsilon^2(\mathbf{x})}{\rho} \end{aligned} \tag{3.10}$$

where the term in the brackets is $\epsilon^2(\mathbf{x})$, which as we have seen can be computed using the first M principal components. We can therefore write the form of the likelihood estimate based on $\hat{d}(\mathbf{x})$ as the product of two marginal and independent Gaussian

[3] We will see shortly that given the typical eigenvalue spectra observed in practice (*e.g.*, Figure 3-1(b)), this choice is optimal for a different reason: it minimizes the information-theoretic *divergence* between the true density and our estimate of it.

densities

$$\hat{P}(\mathbf{x}|\Omega) = \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right]\cdot\left[\frac{\exp\left(-\frac{\epsilon^2(\mathbf{x})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \tag{3.11}$$

$$= P_F(\mathbf{x}|\Omega)\,\hat{P}_{\bar{F}}(\mathbf{x}|\Omega)$$

where $P_F(\mathbf{x}|\Omega)$ is the true marginal density in F-space and $\hat{P}_{\bar{F}}(\mathbf{x}|\Omega)$ is the estimated marginal density in the orthogonal complement $\bar{F}$-space. The optimal value of ρ can now be determined by minimizing a suitable cost function $J(\rho)$. From an information-theoretic point of view, this cost function should be the Kullback-Leibler divergence or *relative entropy* [9] between the true density $P(\mathbf{x}|\Omega)$ and its estimate $\hat{P}(\mathbf{x}|\Omega)$

$$J(\rho) = \int P(\mathbf{x}|\Omega)\,\log\frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)}\,d\mathbf{x} = \mathrm{E}\left[\log\frac{P(\mathbf{x}|\Omega)}{\hat{P}(\mathbf{x}|\Omega)}\right] \tag{3.12}$$

Using the diagonalized forms of the *Mahalanobis* distance $d(\mathbf{x})$ and its estimate $\hat{d}(\mathbf{x})$ and the fact that $\mathrm{E}[y_i^2] = \lambda_i$, it can be easily shown that (see Appendix A)

$$J(\rho) = \frac{1}{2}\sum_{i=M+1}^{N}\left[\frac{\lambda_i}{\rho} - 1 + \log\frac{\rho}{\lambda_i}\right] \tag{3.13}$$

The optimal weight ρ^* can be then found by minimizing this cost function with respect to ρ. Solving the equation $\frac{\partial J}{\partial \rho} = 0$ yields (see Appendix A)

$$\rho^* = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i \tag{3.14}$$

which is simply the arithmetic average of the eigenvalues in the orthogonal subspace $\bar{F}$. In addition to its optimality, ρ^* also results in an *unbiased* estimate of the *Mahalanobis* distance — *i.e,* $\mathrm{E}[\hat{d}(\mathbf{x};\rho^*)] = \mathrm{E}[d(\mathbf{x})]$ (see Appendix A). What this derivation shows is that once we select the M-dimensional principal subspace F (as indicated, for example, by PCA), the optimal estimate of the sufficient statistic $\hat{d}(\mathbf{x})$ has the form of Eq.(3.10) with ρ given by Eq.(3.14).

It is interesting to consider the minimal cost $J(\rho^*)$

$$J(\rho^*) = \frac{1}{2} \sum_{i=M+1}^{N} \log \frac{\rho^*}{\lambda_i} \tag{3.15}$$

from the point of view of the $\bar{F}$-space eigenvalues $\{\lambda_i : i = M+1, \cdots, N\}$. It is easy to show that $J(\rho^*)$ is minimized when the $\bar{F}$-space eigenvalues have the *least* spread about their mean ρ^*. This suggests a strategy for selecting the principal subspace: choose F such that the eigenvalues associated with its orthogonal complement $\bar{F}$ have the least absolute deviation about their mean. In practice, the higher-order eigenvalues typically decay and stabilize near the observation noise variance. Therefore this strategy is usually consistent with the standard PCA practice of discarding the higher-order components since these tend to correspond to the "flattest" portion of the eigenvalue spectrum (see Figure 3-1(b)). In the limit, as the $\bar{F}$-space eigenvalues become exactly equal, the divergence $J(\rho^*)$ will be zero and our density estimate $\hat{P}(\mathbf{x}|\Omega)$ approaches the true density $P(\mathbf{x}|\Omega)$.

We note that in most applications it is customary to simply discard the $\bar{F}$-space component and simply work with $P_F(\mathbf{x}|\Omega)$. However, the use of the DFFS metric or equivalently the marginal density $P_{\bar{F}}(\mathbf{x}|\Omega)$ is critically important in formulating the likelihood of an observation $\mathbf{x}$ — especially in an object detection task — since there are an infinity of vectors which are *not* members of Ω which can have likely F-space projections. Without $P_{\bar{F}}(\mathbf{x}|\Omega)$ a detection system can result in a significant number of false alarms.

3.3 Multimodal Densities

In the previous section we assumed that the probability density of the training images was Gaussian. This lead to a likelihood estimate in the form of a product of two independent multivariate Gaussian distributions (or equivalently the sum of two *Mahalanobis* distances: DIFS + DFFS). In our experience, the distribution of samples in the feature space is often accurately modeled by a single Gaussian distribution. This

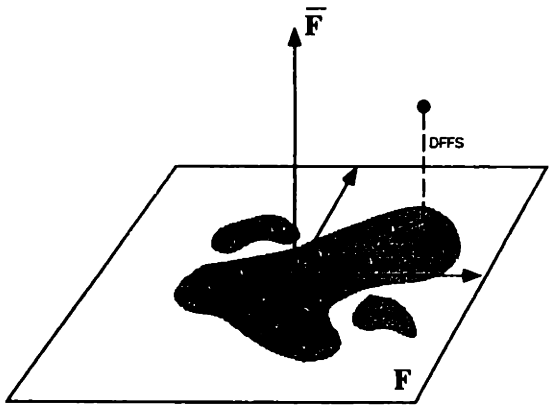


Figure 3-2: Decomposition into the principal subspace F and its orthogonal complement $\bar{F}$ for an arbitrary density.

is especially true in cases where the training images are accurately aligned views of similar objects seen from a standard view (*e.g.*, aligned frontal views of human faces at the same scale and lighting conditions). However, when the training set represents multiple views or multiple objects under varying illumination conditions, the distribution of training views in F-space is no longer unimodal. In fact the training data tends to lie on complex and non-separable low-dimensional manifolds in image space. One way to tackle this multimodality is to build a view-based (or object-based) formulation where separate eigenspaces are used for each view [32]. Another approach is to capture the complexity of these manifolds in a universal or *parametric* eigenspace using splines [25], or local basis functions [5].

If we assume that the $\bar{F}$-space components are Gaussian and independent of the principal features in F (this would be true in the case of pure observation noise in $\bar{F}$) we can still use the separable form of the density estimate $\hat{P}(\mathbf{x}|\Omega)$ in Eq.(3.11) where $P_F(\mathbf{x}|\Omega)$ is now an *arbitrary* density $P(\mathbf{y})$ in the principal component vector $\mathbf{y}$. Figure 3-2 illustrates the decomposition, where the DFFS is the residual $\epsilon^2(\mathbf{x})$ as before. The DIFS, however, is no longer a simple *Mahalanobis* distance but can nevertheless be interpreted as a "distance" by relating it to $P(\mathbf{y})$ — *e.g.*, as DIFS $= -\log P(\mathbf{y})$.

The density $P(\mathbf{y})$ can be estimated using a parametric mixture model. Specifically,

we can model arbitrarily complex densities using a Mixture-of-Gaussians

$$P(\mathbf{y}|\Theta) = \sum_{i=1}^{N_c} \pi_i \, g(\mathbf{y}; \mu_i, \Sigma_i) \tag{3.16}$$

where $g(\mathbf{y}; \mu, \Sigma)$ is an M-dimensional Gaussian density with mean vector μ and covariance Σ, and the π_i are the mixing parameters of the components, satisfying $\sum \pi_i = 1$. The mixture is completely specified by the parameter $\Theta = \{\pi_i, \mu_i, \Sigma_i\}_{i=1}^{N_c}$. Given a training set $\{\mathbf{y}^t\}_{t=1}^{N_T}$ the mixture parameters can be estimated using the ML principle

$$\Theta^* = \text{argmax} \left[\prod_{t=1}^{N_T} P(\mathbf{y}^t|\Theta) \right] \tag{3.17}$$

This estimation problem is best solved using the Expectation-Maximization (EM) algorithm [12] which consists of the following two-step iterative procedure:

- E-step:

$$h_i^k(t) = \frac{\pi_i^k g(\mathbf{y}^t; \mu_i^k, \Sigma_i^k)}{\sum_{j=1}^{N_c} \pi_j^k g(\mathbf{y}^t; \mu_j^k, \Sigma_i^k)} \tag{3.18}$$

- M-step:

$$\pi_i^{k+1} = \frac{\sum_{t=1}^{N_T} h_i^k(t)}{\sum_{i=1}^{N_c} \sum_{t=1}^{N_T} h_i^k(t)} \tag{3.19}$$

$$\mu_i^{k+1} = \frac{\sum_{t=1}^{N_T} h_i^k(t) \mathbf{y}^t}{\sum_{t=1}^{N_T} h_i^k(t)} \tag{3.20}$$

$$\Sigma_i^{k+1} = \frac{\sum_{t=1}^{N_T} h_i^k(t)(\mathbf{y}^t - \mu_i^{k+1})(\mathbf{y}^t - \mu_i^{k+1})^T}{\sum_{t=1}^{N_T} h_i^k(t)} \tag{3.21}$$

The E-step computes the *a posteriori* probabilities $h_i(t)$ which are the *expectations* of "missing" component labels $z_i(t) = \{0, 1\}$ which denote the membership of $\mathbf{y}^t$

in the i-th component. Once these expectations have been computed, the M-step maximizes the joint likelihood of the data *and* the "missing" variables $z_i(t)$. The EM algorithm is monotonically convergent in *likelihood* and is thus guaranteed to find a local maximum in the total likelihood of the training set. Further details of the EM algorithm for estimation of mixture densities can be found in [39].

Given our operating assumptions — that the training data is M-dimensional (at most) and resides solely in the principal subspace F with the exception of perturbations due to white Gaussian measurement noise, or equivalently that the $\bar{F}$-space component of the data is itself a separable Gaussian density — the estimate of the complete likelihood function $P(\mathbf{x}|\Omega)$ is given by

$$\hat{P}(\mathbf{x}|\Omega) \;=\; P(\mathbf{y}|\Theta^*)\;\hat{P}_F(\mathbf{x}|\Omega) \tag{3.22}$$

where $\hat{P}_F(\mathbf{x}|\Omega)$ is a Gaussian component density based on the DFFS, as before.

Chapter 4

Probabilistic Detection

In this chapter we examine the use of the eigenspace density estimates derived in the previous chapter for visual object detection of frontal faces, facial features and hands.

4.1 Maximum Likelihood Detection

The density estimate $\hat{P}(\mathbf{x}|\Omega)$ can be used to compute a local measure of target saliency at each spatial position (i,j) in an input image based on the vector $\mathbf{x}$ obtained by the lexicographic ordering of the pixel values in a local neighborhood R

$$S(i,j;\Omega) = \hat{P}(\mathbf{x}|\Omega)\ , \quad \mathbf{x} = \downarrow[\{I(i+r,j+c) : (r,c) \in R\}] \tag{4.1}$$

where $\downarrow[\bullet]$ is the operator which converts a subimage into a vector. The ML estimate of position of the target Ω is then given by

$$(i,j)^{\mathrm{ML}} = \operatorname{argmax} S(i,j;\Omega) \tag{4.2}$$

This is illustrated in Figure 4-1.

This ML formulation can be extended to estimate object scale with *multiscale* saliency maps. The likelihood computation is performed (in parallel) on linearly scaled versions of the input image $I^{(\sigma)}$ corresponding to a pre-determined set of scales

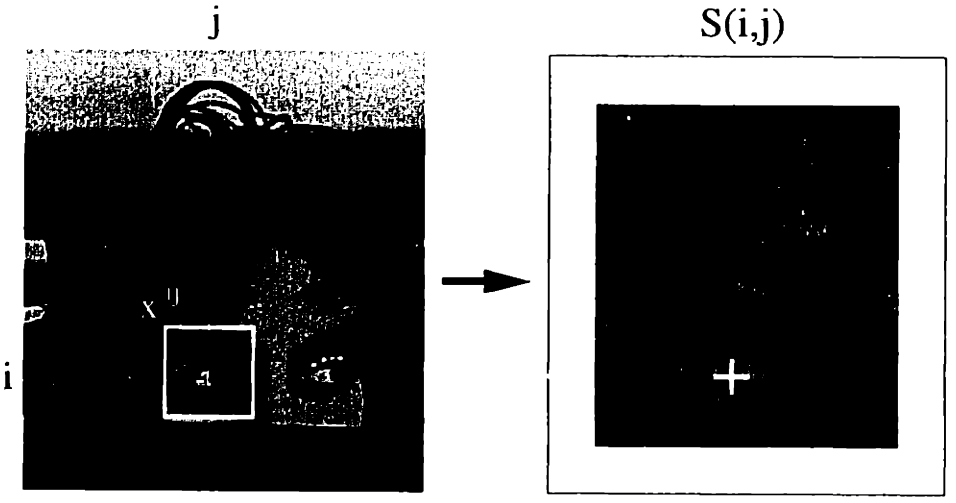


Figure 4-1: Target saliency map S(i,j) showing the probability of a left eye pattern over the input image.

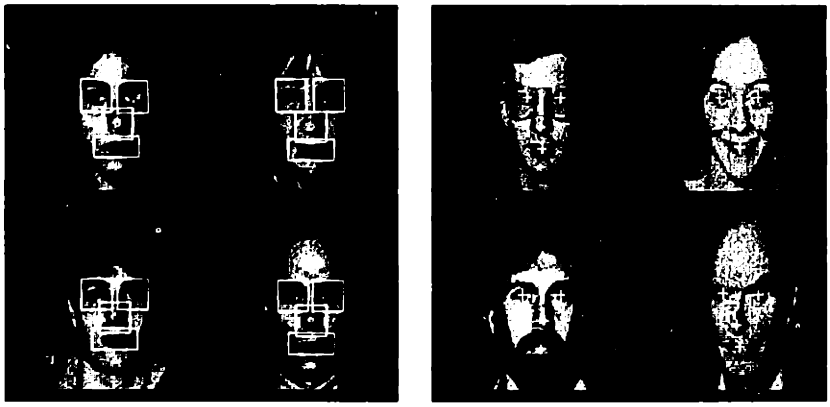
$\{\sigma_1, \sigma_2, \cdots \sigma_n\}$

$$S(i,j,k;\ \Omega) = \hat{P}\left(\downarrow \{I^{(\sigma_k)}(\sigma_k i + r, \sigma_k j + c) : (r,c) \in R\} \mid \Omega\right) \tag{4.3}$$

where the ML estimate of the spatial and scale indices is defined by

$$(i,j,k)^{\mathrm{ML}} \ = \ \mathrm{argmax}\ S(i,j,k\ ;\ \Omega) \tag{4.4}$$

One important factor of variability in the appearance of the object in grayscale imagery is that of lighting and contrast. While compensation for variable lighting direction is difficult, one can normalize for global (ambient) illumination changes (as well as the linear response characteristics of the CCD camera) by normalizing each subimage **x** by its mean and standard deviation. This contrast normalization is performed both during training (density estimation) and also in the operational mode (*e.g.*, in detection).



(a) (b)

Figure 4-2: (a) Examples of facial feature training templates and (b) the resulting typical detections.

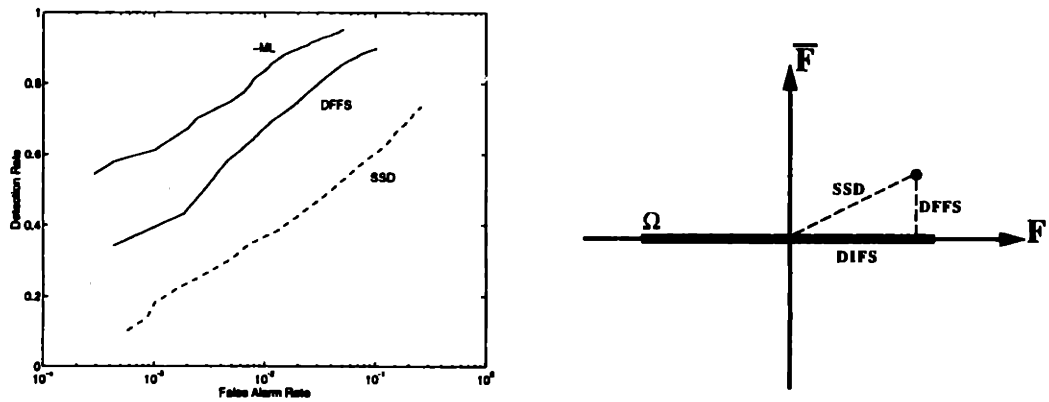
4.2 Applications

The above ML detection technique has been tested in the detection of complex natural objects including human faces, facial features (*e.g.*, eyes), as well as non-rigid and articulated objects such as human hands. In this section we will present several examples from these application domains.

4.2.1 Faces

Over the years, various strategies for facial feature detection have been proposed, ranging from edge-map projections [18], to more recent techniques using generalized symmetry operators [40] and multilayer perceptrons [45]. In any robust face processing system this task is critically important since a face must be first geometrically normalized by aligning its features with those of a stored model before recognition can be attempted.

The eigentemplate approach to the detection of facial features in "mugshots" was proposed in [32], where the DFFS metric was shown to be superior to standard template matching for target detection. The detection task was the estimation of the position of facial features (the left and right eyes, the tip of the nose and the center of the mouth) in frontal view photographs of faces at fixed scale. Figure 4-2 shows



(a) (b)

Figure 4-3: (a) Detection performance of an SSD, DFFS and a ML detector, (b) geometric interpretation of the detectors.

examples of facial feature training templates and the resulting detections on the MIT Media Laboratory's database of 7,562 "mugshots".

We have compared the detection performance of three different detectors on approximately 7,000 test images from this database: a sum-of-square-differences (SSD) detector based on the average facial feature (in this case the left eye), an eigentemplate or DFFS detector and a ML detector based on $S(i,j;\Omega)$ as defined in section 3.2. Figure 4-3(a) shows the *receiver operating characteristic* (ROC) curves for these detectors, obtained by varying the detection threshold independently for each detector. The DFFS and ML detectors were computed based on a 5-dimensional principal subspace. Since the projection coefficients were unimodal a Gaussian distribution was used in modeling the true distribution for the ML detector as in section 3.2. Note that the ML detector exhibits the best detection vs. false-alarm tradeoff and yields the highest detection rate (of 95%). Indeed, at the *same* detection rate the ML detector has a false-alarm rate which is nearly 2 orders of magnitude lower than the SSD.

Figure 4-3(b) provides the geometric intuition regarding the operation of these detectors. The SSD detector's threshold is based on the *radial* distance between the average template (the origin of this space) and the input pattern. This leads to hyperspherical detection regions about the origin. In contrast, the DFFS detector measures the orthogonal distance to F, thus forming planar acceptance regions about

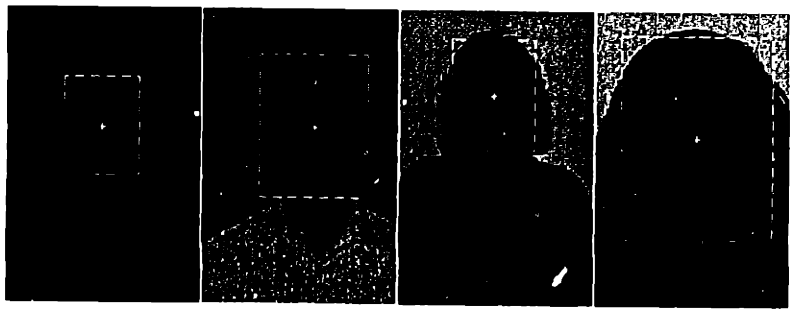


Figure 4-4: Multiscale Face Detection

F. Consequently to accept valid object patterns in Ω which are very different from the mean, the SSD detector must operate with high thresholds which consequently lead to many false alarms. But at the same time, the DFFS detector can not discriminate between the object class Ω and non-Ω patterns in F. The solution is provided by the ML detector which incorporates both the $\bar{F}$-space component (DFFS) and the F-space likelihood (DIFS). The probabilistic interpretation of Figure 4-3(b) is as follows: SSD assumes a *single* prototype (the mean) in additive white Gaussian noise whereas the DFFS assumes a *uniform* density in F. The ML detector, on the other hand, uses the complete probability density for detection.

We have incorporated and tested the multiscale version of the ML detection technique in a face detection task. This multiscale head finder was tested on the ARPA FERET database where 308 out of 310 faces were correctly detected. Figure 4-4 shows examples of the ML estimate of the position and scale on these images. The multiscale saliency maps $S(i, j, k; \Omega)$ were computed based on the likelihood estimate $\hat{P}(\mathbf{x}|\Omega)$ in a 10-dimensional principal subspace using a Gaussian model (section 3.2). Note that this detector is able to localize the position and scale of the head despite variations in hair style and hair color, as well as presence of sunglasses. Illumination invariance was obtained by normalizing the input subimage $\mathbf{x}$ to a zero-mean unit-norm vector.

We have also used the multiscale version of the ML detector as the *attentional* component of an automatic system for recognition and model-based coding of faces. The block diagram of this system is shown in Figure 5-5 which consists of a two-stage

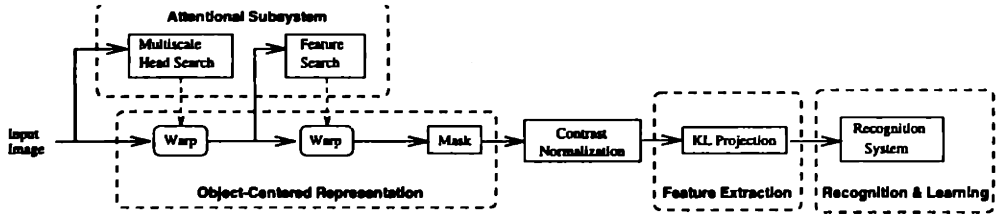
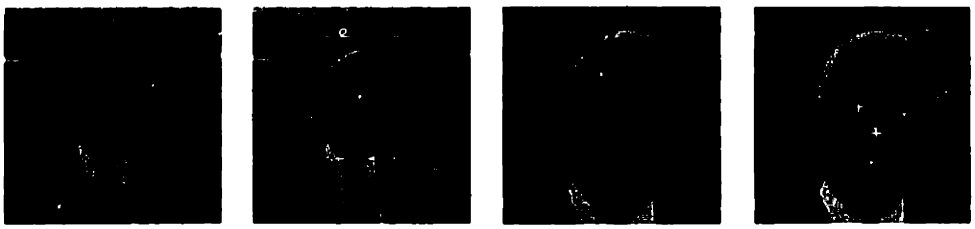


Figure 4-5: The face processing system.

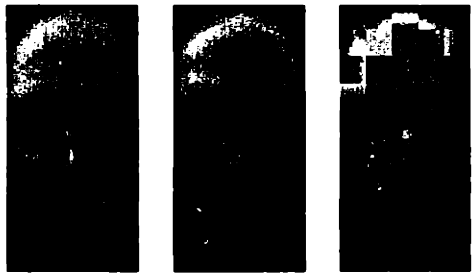


(a) (b) (c) (d)

Figure 4-6: (a) original image, (b) position and scale estimate, (c) normalized head image, (d) position of facial features.

object detection and alignment stage, a contrast normalization stage, and a feature extraction stage whose output is used for both recognition and coding. Figure 4-6 illustrates the operation of the detection and alignment stage on a natural test image containing a human face. The function of the face finder is to locate regions in the image which have a high likelihood of containing a face.

The first step in this process is illustrated in Figure 4-6(b) where the ML estimate of the position and scale of the face are indicated by the cross-hairs and bounding box. Once these regions have been identified, the estimated scale and position are used to normalize for translation and scale, yielding a standard "head-in-the-box" format image (Figure 4-6(c)). A second feature detection stage operates at this fixed scale to estimate the position of 4 facial features: the left and right eyes, the tip of the nose and the center of the mouth (Figure 4-6(d)). Once the facial features have been detected, the face image is warped to align the geometry and shape of the face with that of a canonical model. Then the facial region is extracted (by applying a fixed mask) and subsequently normalized for contrast. The geometrically aligned and



(a) (b) (c)

Figure 4-7: (a) aligned face, (b) eigenspace reconstruction (85 bytes) (c) JPEG reconstruction (530 bytes).

Figure 4-8: The first 8 eigenfaces.

normalized image (shown in Figure 4-7(a)) is then projected onto a custom set of eigenfaces to obtain a feature vector which is then used for recognition purposes as well as facial image coding.

Figure 4-7 shows the normalized facial image extracted from Figure 4-6(d), its reconstruction using a 100-dimensional eigenspace representation (requiring only 85 bytes to encode) and a comparable non-parametric reconstruction obtained using a standard transform-coding approach for image compression (requiring 530 bytes to encode). This example illustrates that the eigenface representation used for recognition is also an effective *model-based* representation for data compression. The first 8 eigenfaces used for this representation are shown in Figure 4-8.

Figure 4-9 shows the results of a similarity search in an image database tool called Photobook [33]. Each face in the database was automatically detected and aligned by the face processing system in Figure 4-5. The normalized faces were then projected onto a 100-dimensional eigenspace. The image in the upper left is the one searched

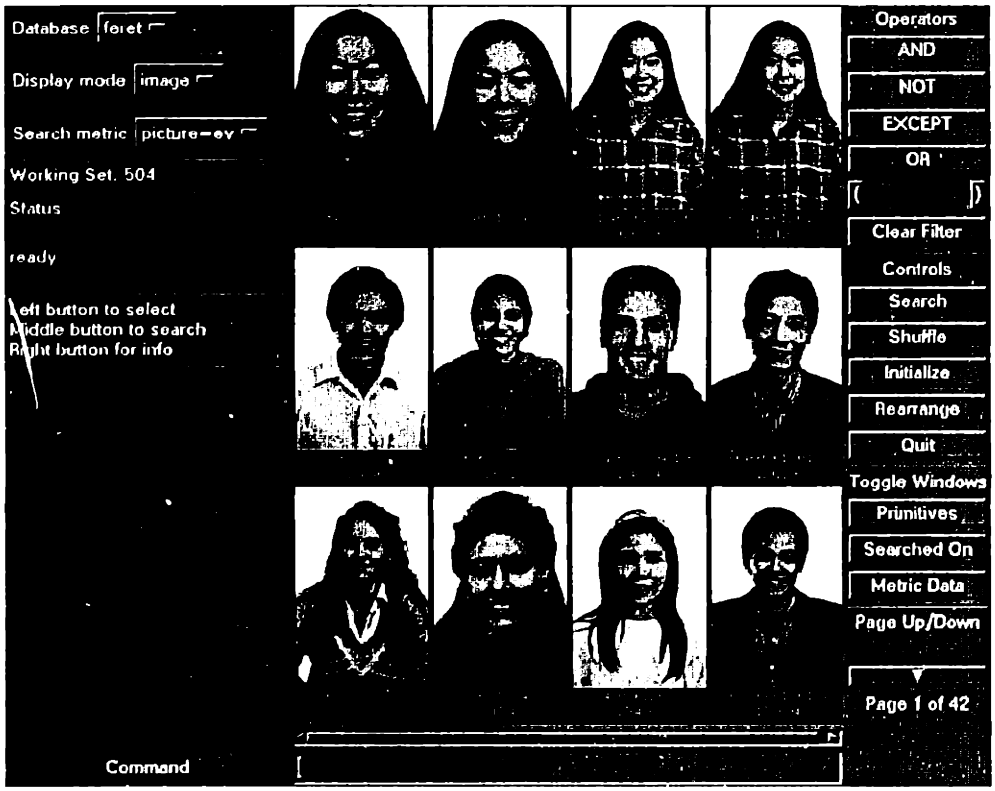


Figure 4-9: Photobook: FERET face database.

on and the remainder are the ranked nearest neighbors in the FERET database. The top three matches in this case are images of the same person taken a month apart and at different scales. The recognition accuracy (defined as the percent correct rank-one matches) on a database of 155 individuals is 99% [24].

4.2.2 Hands

We have also applied our eigenspace density estimation technique to articulated and non-rigid objects such as hands. In this particular domain, however, the original intensity image is an unsuitable representation since, unlike faces, hands are essentially textureless objects. Their identity is characterized by the variety of *shapes* they can assume. For this reason we have chosen an edge-based representation of hand shapes which is invariant with respect to illumination, contrast and scene background. A training set of hand gestures was obtained against a black background. The 2D con-

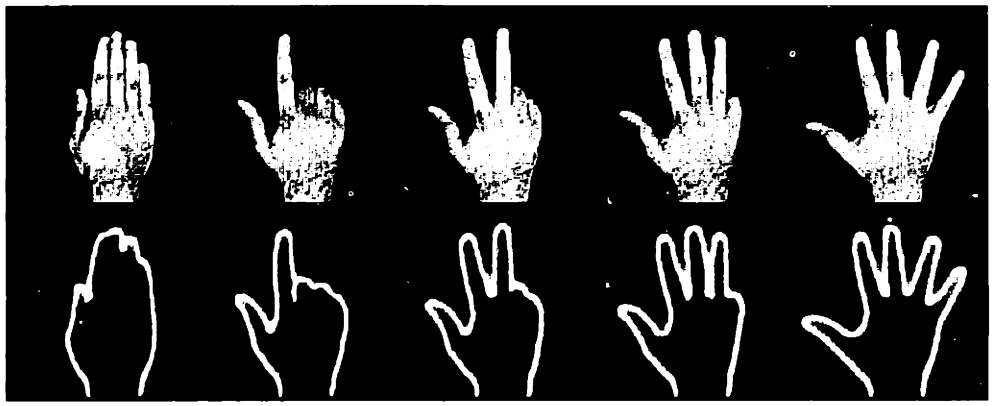
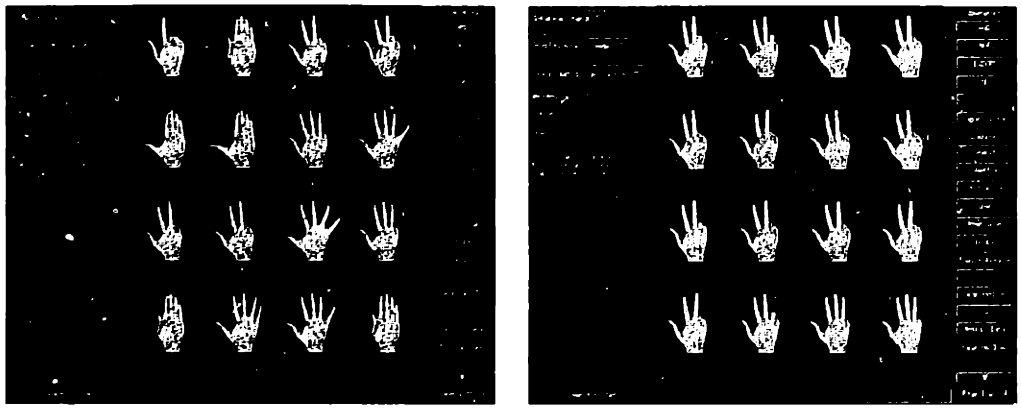


Figure 4-10: Examples of hand gestures and their diffused edge representation.

tour of the hand was then extracted using a Canny edge-operator. These binary edge maps, however, are highly uncorrelated with each other due to their sparse nature. This leads to a very high-dimensional principal subspace. Therefore to reduce the intrinsic dimensionality, we *induced* spatial correlation via a *diffusion* process on the binary edge map, which effectively broadens and "smears" the edges, yielding a continuous-valued contour image which represents the object shape in terms of the spatial distribution of edges. Figure 4-10 shows examples of training images and their diffused edge map representations. Note that this *spatiotopic* representation of shape is interesting because it is consonant with our knowledge of biological representations, especially as compared to approaches motivated purely by computational considerations (*e.g.*, moments [15], Fourier descriptors [36], "snakes" [19], Point Distribution Models (PDM) [8], and modal descriptions [42]).

It is important to verify whether such a representation is valid for modeling hand shapes. Therefore we tested the diffused contour image representation in a recognition experiment which yielded a 100% rank-one accuracy on 375 frames from an image sequence containing 7 hand gestures. The matching technique was a nearest-neighbor classification rule in a 16-dimensional principal subspace. Figure 4-11(a) shows some examples of the various hand gestures used in this experiment. Figure 4-11(b) shows the 15 images that are most similar to the "two" gesture appearing in the top left. Note that the hand gestures judged most similar are all objectively the same gesture.



(a) (b)

Figure 4-11: (a) a random assortment of hand gestures (b) images ordered by similarity (left-to-right, top-to-bottom) to the image at the upper left.

Naturally, the success of such a recognition system is critically dependent on the ability to find the hand (in any of its articulated states) in a cluttered scene, to account for its scale and to align it with respect to an object-centered reference frame prior to recognition. This localization was achieved with the same multiscale ML detection paradigm used with faces, with the exception that the underlying image representation of the hands was the diffused edge map rather the grayscale image.

The probability distribution of hand shapes in this representation was automatically learned using our eigenspace density estimation technique. In this case, however, the distribution of training data is *multimodal* due to the different hand shapes for each gesture. Therefore the multimodal density estimation technique in section 3.3 was used. Figure 4-12(a) shows a projection of the training data on the first two dimensions of the principal subspace F (defined in this case by $M = 16$) which exhibit the underlying multimodality of the data. Figure 4-12(b) shows a 10-component Mixture-of-Gaussians density estimate for the training data. The parameters of this estimate were obtained with 20 iterations of the EM algorithm. The orthogonal $\bar{F}$-space component of the density was modeled with a Gaussian distribution as in section 3.3.

The resulting complete density estimate $\hat{P}(\mathbf{x}|\Omega)$ was then used in a detection experiment on test imagery of hand gestures against a cluttered background scene.

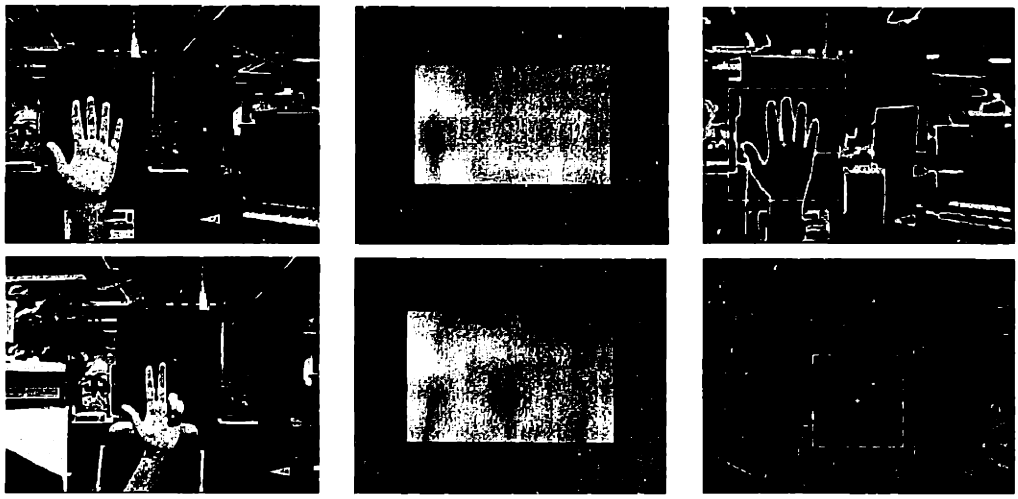
(a) (b)

Figure 4-12: (a) Distribution of training hand shapes (shown in the 1st two dimensions of the principal subspace) (b) Mixture-of-Gaussians fit using 10 components.

In accordance with our representation, the input imagery was first pre-processed to generate a diffused edge map and then scaled accordingly for a multiscale saliency computation. Figure 4-13 shows two examples from the test sequence, where we have shown the original image, the negative log-likelihood saliency map, and the ML estimates of position and scale (superimposed on the diffused edge map). Note that these examples represent two different hand gestures at slightly different scales. We note that the success of this detection scheme is dependent on a visible and mostly complete hand contour in the edge map, which places some restrictions on the imaging situation and background.

To better quantify the performance of the ML detector on hands we carried out the following experiment. The original 375-frame video sequence of training hand gestures was divided into 2 parts. The first (training) half of this sequence was used for learning, including computation of the KL basis and the subsequent EM clustering. For this experiment we used a 5-component mixture in a 10-dimensional principal subspace. The second (testing) half of the sequence was then embedded in the background scene, which contains a variety of shapes. In addition, severe noise conditions were simulated as shown in Figure 4-14(a).

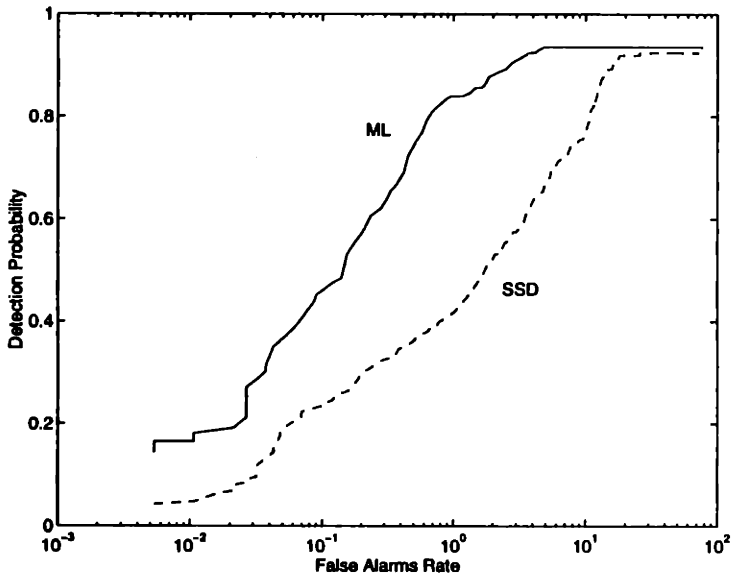
We then compared the detection performance of an SSD detector (based on the mean edge-based hand representation) and a probabilistic detector based on the com-



(a) (b) (c)

Figure 4-13: (a) Original grayscale image, (b) negative log-likelihood map (at most likely scale) and (c) ML estimate of position and scale superimposed on edge-map.

plete estimated density. The resulting negative-log-likelihood detection maps were passed through a valley-detector to isolate local minimum candidates which were then subjected to a ROC analysis. A correct detection was defined as a below-threshold local minimum within a 5-pixel radius of the ground truth target location. Figure 4-14(b) shows the performance curves obtained for the two detectors We note, for example, that at an 85% detection probability the ML detector yields (on the average) 1 false alarm per frame, where as the SSD detector yields an order of magnitude more false alarms.



(a) (b)

Figure 4-14: (a) Example of test frame containing a hand gesture amidst severe background clutter and (b) ROC curve performance contrasting SSD and ML detectors.

Chapter 5

Probabilistic Recognition

In the previous chapter we demonstrated how the subspace density estimates can be used for target detection. Now we will see how they also can be used for visual recognition. From a probabilistic perspective, the class conditional density $P(\mathbf{x}|\Omega)$ is the most important object representation to be learned. This density is the critical component in detection, recognition, prediction, interpolation and general inference. For example, having learned these densities for several object classes $\{\Omega_1, \Omega_2, \cdots, \Omega_n\}$, one can invoke a Bayesian framework for classification and recognition:

$$P(\Omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega_i)P(\Omega_i)}{\sum_{j=1}^{n} P(\mathbf{x}|\Omega_j)P(\Omega_j)} \tag{5.1}$$

where now a maximum *a posteriori* (MAP) classification rule can be used for object/pose identification.

One disadvantage of this technique is that it would require many views to estimate individual densities, requiring multiple images for each object. A more computationally attractive alternative is to redefine the similarity measure such that recognition can be performed from a single view. In this chapter we introduce a probabilistic similarity measure which has this characteristic.

5.1 Similarity Measures

Current approaches to image matching for visual object recognition and image database retrieval often make use of simple image similarity metrics such as Euclidean distance or normalized correlation, which correspond to a standard template-matching approach to recognition. For example, in its simplest form, the similarity measure $S(I_1, I_2)$ between two images I_1 and I_2 can be set to be inversely proportional to the norm $||I_2 - I_1||$. Such a simple formulation suffers from two major drawbacks: it requires precise alignment of the objects in the image and does not exploit knowledge of which type of variations are critical (as opposed to incidental) in expressing similarity. In this chapter, we formulate a *probabilistic* similarity measure which is based on the probability that the image-based differences, denoted by $d(I_1, I_2)$, are characteristic of typical variations in appearance of the *same* object. For example, for purposes of face recognition, we can define two classes of facial image variations: *intrapersonal* variations Ω_I (corresponding, for example, to different facial expressions of the *same* individual) and *extrapersonal* variations Ω_E (corresponding to variations between *different* individuals). Our similarity measure is then expressed in terms of the probability

$$S(I_1, I_2) = P(d(I_1, I_2) \in \Omega_I) = P(\Omega_I | d(I_1, I_2)) \tag{5.2}$$

where $P(\Omega_I | d(I_1, I_2))$ is the *a posteriori* probability given by Bayes rule, using estimates of the likelihoods $P(d(I_1, I_2)|\Omega_I)$ and $P(d(I_1, I_2)|\Omega_E)$ which are derived from training data using the subspace method for density estimation of high-dimensional data developed in Chapter 3.

5.2 Representations for $d(I_1, I_2)$

Furthermore, we use a novel representation for $d(I_1, I_2)$ which combines both the spatial (XY) and grayscale (I) components of the image in a unified XYI framework (unlike previous approaches which essentially treat the shape and texture components

independently, *e.g.*, [10, 11, 22, 3]). Specifically, I_1 is modeled as a physically-based deformable 3D surface (or manifold) in XYI-space which deforms in accordance with attractive "physical forces" exerted by I_2. The dynamics of this system are efficiently solved for using the *analytic modes of vibration* [26], yielding a 3D correspondence field for warping I_1 into I_2. In addition, we use the *parametric* representation, $d(I_1, I_2) = \tilde{\mathbf{U}}$, where $\tilde{\mathbf{U}}$ is the modal amplitude spectrum of the resultant deformation [28]. This manifold matching technique can be viewed as a more general formulation for image correspondence which, unlike optical flow, does *not* require a constant brightness assumption [14]. In fact, by simply disabling the I component of our deformations we can obtain a standard 2D deformable mesh which yields correspondences similar to an optical flow technique with thin-plate regularizers.

Finally, we experimentally compare our deformable matching technique with two alternative (non-deformable) methods: one using intensity differences with

$$d(I_1, I_2) = I_2 - I_1$$

and a standard correspondence method using optical flow with

$$d(I_1, I_2) = flow(I_1, I_2)$$

where $flow(I_1, I_2)$ is the vector flow field between I_1 and I_2. We note that these two methods can be viewed as degenerate cases of our general XYI correspondence method: the former assumes XY correspondences and makes the I difference explicit, whereas the latter assumes comparable I components and makes the XY variations explicit. Our experimental results have confirmed our basic intuition that the fully deformable XYI warping method yields the best characterization of $d(I_1, I_2)$, at least as far as recognition is concerned. The advantage of our method over optical flow is key, since this simpler method relies all too heavily on the constant brightness assumption and is prone to failure when there are large grayscale variations between the images of different individuals (*e.g.*, presence/absence of facial hair).

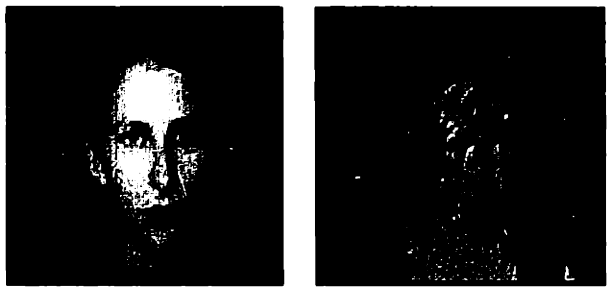


Figure 5-1: An image and its XYI surface representation

5.3 XYI Image Warping

As shown in [28], we can formulate an image matching technique based on a 3D surface representation of an image $I(x,y)$ — *i.e.,* as the surface $(x,y,I(x,y))$ as shown, for example, in Figure 5-1 -- and developed an efficient method to *warp* one image onto another using a physically-based deformation model. In this section we briefly review the mathematics of this approach (for further details the reader is referred to [27, 28]).

The intensity surface is modeled as a deformable mesh and is governed by Lagrangian dynamics [2] :

$$\mathbf{M}\ddot{\mathbf{U}} + \mathbf{C}\dot{\mathbf{U}} + \mathbf{K}\mathbf{U} = \mathbf{F}(t) \tag{5.3}$$

where $\mathbf{U} = [\ldots, \Delta x_i, \Delta y_i, \Delta z_i, \ldots]^T$ is a vector storing nodal displacements, $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ are respectively the mass, damping and stiffness matrices of the system, and $\mathbf{F}$ is the external force. In warping one image onto a second (reference) image, the external force at each node M_i of the mesh points is the vector to the closest 3D point P_i in the reference surface:

$$\mathbf{F}(t) = [\ldots, \overrightarrow{M_i P_i}(t), \ldots]^T \tag{5.4}$$

The final correspondence (and consequently the resultant XYI-warp) between two images is obtained by solving the governing equation above. Figure 5-2 shows a schematic representation of the deformation process. Note that the external forces (dashed arrows) do *not* necessarily correspond to the final displacement field of the

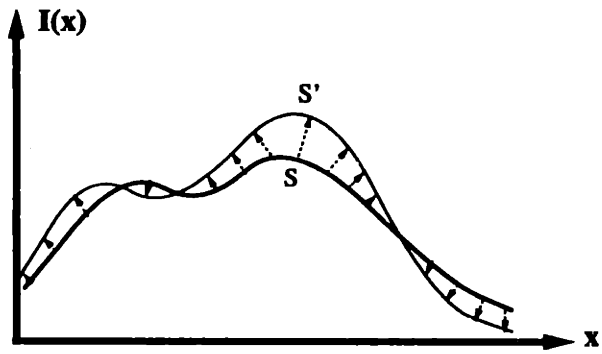


Figure 5-2: A cross-section of the intensity surface S being pulled towards S' by image forces

surface. The elasticity of the surface provides an intrinsic smoothness constraint for computing the final displacement field.

We note that this formulation provides an interesting alternative to optical flow methods for obtaining correspondence, without the classical *brightness constraint* [14]. Indeed, the brightness constraint corresponds to a particular case of our formulation where the closest point P_i has to have the same intensity as M_i — *i.e.*, $\overrightarrow{M_iP_i}$ is parallel to the XY plane. We do not make that assumption here.

Solutions of the governing equation are typically obtained using an eigenvector-based *modal* decomposition [35, 27, 26]. In particular, the vibration modes $\phi(i)$ of the previous deformable surface are the vector solutions of the eigenproblem :

$$\mathbf{K}\boldsymbol{\phi} = \omega^2\mathbf{M}\boldsymbol{\phi} \tag{5.5}$$

where $\omega(i)$ is the i-th eigenfrequency of the system. Solving the governing equations in the modal basis leads to scalar equations where the unknown $\tilde{u}(i)$ is the amplitude of mode i [2]

$$\ddot{\tilde{u}}(i) + \tilde{c}_i\dot{\tilde{u}}(i) + \omega(i)^2\tilde{u}(i) = \tilde{f}_i(t) \qquad i = 1, \ldots, 3N. \tag{5.6}$$

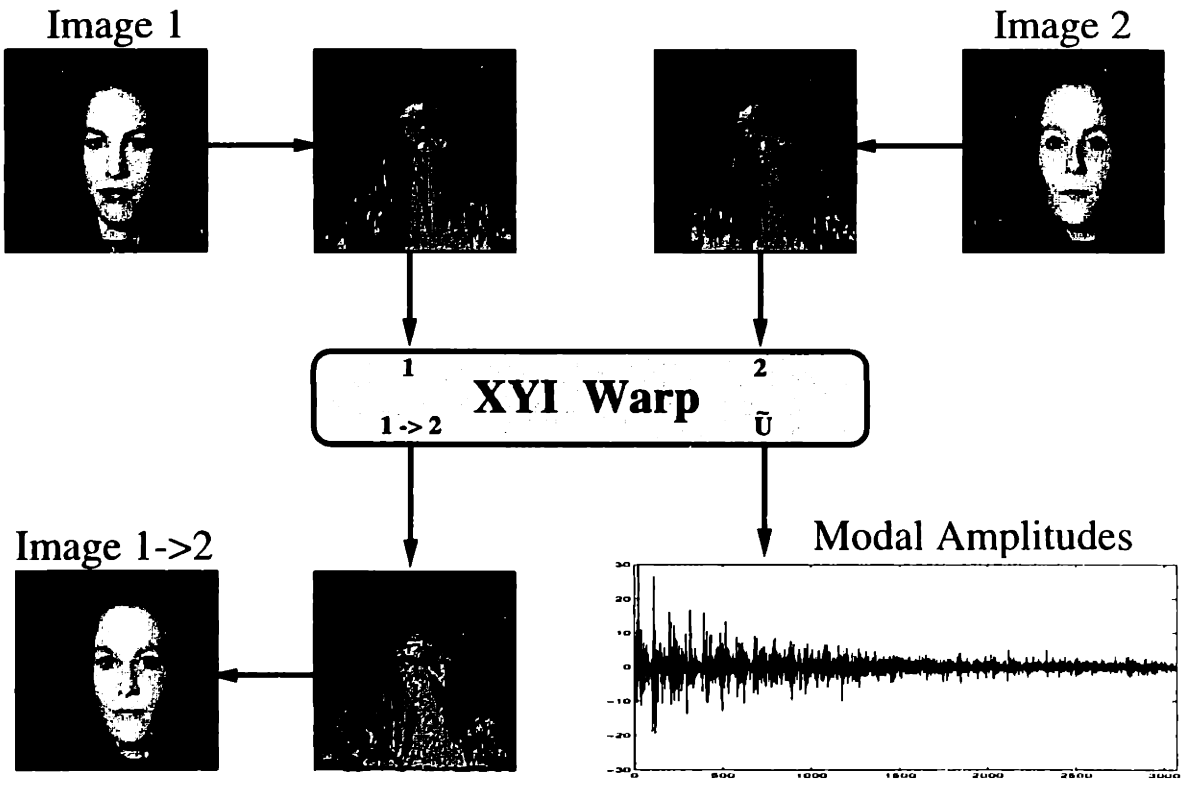


Figure 5-3: Example of XYI warping two images.

The closed-form expression of the displacement field is then given by

$$\mathbf{U} \approx \sum_{i=1}^{P} \check{u}(i)\boldsymbol{\phi}(i) \tag{5.7}$$

with $P \ll 3N$, which means that only P scalar equations of the type of (5.6) need to be solved. The modal superposition equation (5.7) can be seen as a Fourier expansion with high-frequencies neglected [26]. In our formulation, however, we make use of the *analytic modes* [26, 28], which are known sine and cosine functions for specific surface topologies

$$\boldsymbol{\phi}(p,p') = [\ldots, \cos\frac{p\pi(2i-1)}{2n}\cos\frac{p'\pi(2j-1)}{2n'}, \ldots]^T \tag{5.8}$$

These analytic expressions avoid costly eigenvector decompositions and furthermore allow the total number of modes to be easily adjusted for the application.

The above modal analysis technique represents a coordinate transform from the

nodal displacement space to the modal amplitude subspace:

$$\tilde{\mathbf{U}} = \mathbf{\Phi}^T \mathbf{U} \tag{5.9}$$

where $\mathbf{\Phi}$ is the matrix of analytic modes $\phi(p, p')$ and $\tilde{\mathbf{U}}$ is the resultant vector of modal amplitudes which encodes the type of deformations which characterize the difference between the two images. In addition, once we have solved for the resultant 3D displacement field we can then warp the original image onto the second in the XYI space and then render a resultant 2D image using simple computer graphics techniques. Figure 5-3 shows an example illustrating this warping process. We note that the warped image $I_{1\to 2}$ is only an incidental by-product of our correspondence method. Since our main goal is image matching we are primarily interested in the modal amplitude spectrum $\tilde{\mathbf{U}}$ for expressing $d(I_1, I_2)$.

5.4 Analysis of Deformations

We now consider the problem of characterizing the type of deformations which occur when matching two images in a face recognition task. We define two distinct and mutually exclusive classes: Ω_I representing *intrapersonal* variations between multiple images of the same individual (*e.g.*, with different expressions and lighting conditions), and Ω_E representing *extrapersonal* variations which result when matching two different individuals. We will assume that both classes are Gaussian-distributed and seek to obtain estimates of the likelihood functions $P(\tilde{\mathbf{U}}|\Omega_I)$ and $P(\tilde{\mathbf{U}}|\Omega_E)$ for a given deformation's modal amplitude vector $\tilde{\mathbf{U}}$.

Given these likelihoods we can define the similarity score $S(I_1, I_2)$ between a pair of images directly in terms of the intrapersonal *a posteriori* probability as given by Bayes rule:

$$\begin{aligned} S(I_1, I_2) &= P(\Omega_I|\tilde{\mathbf{U}}) \\ &= \frac{P(\tilde{\mathbf{U}}|\Omega_I)P(\Omega_I)}{P(\tilde{\mathbf{U}}|\Omega_I)P(\Omega_I) + P(\tilde{\mathbf{U}}|\Omega_E)P(\Omega_E)} \end{aligned} \tag{5.10}$$

where the priors $P(\Omega)$ can be set to reflect specific operating conditions (*e.g.*, num-

ber of test images *vs.* the size of the database) or other sources of *a priori* knowledge regarding the two images being matched. Additionally, this particular Bayesian formulation casts the standard face recognition task (essentially an M-ary classification problem for M individuals) into a *binary* pattern classification problem with Ω_I and Ω_E. This simpler problem is then solved using the maximum *a posteriori* (MAP) rule — *i.e.,* two images are determined to belong to the same individual if $P(\Omega_I|\tilde{\mathbf{U}}) > P(\Omega_E|\tilde{\mathbf{U}})$, or equivalently, if $S(I_1, I_2) > \frac{1}{2}$.

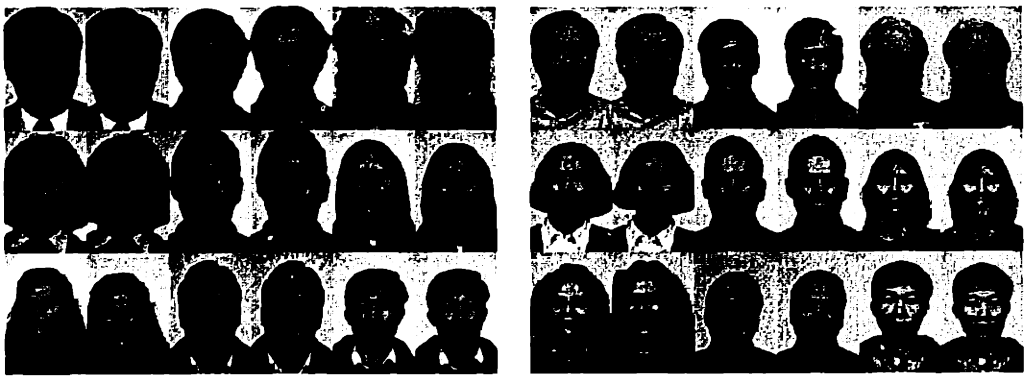
5.4.1 Statistical Modeling of Modes

One difficulty with this approach is that the modal amplitude vectors are high-dimensional, with $\tilde{\mathbf{U}} \in \mathcal{R}^N$ with $N = O(10^3)$. Therefore we typically lack sufficient independent training observations to compute reliable 2nd-order statistics for the likelihood densities (*i.e.,* singular covariance matrices will result). Even if we were able to estimate these statistics, the computational cost of evaluating the likelihoods is formidable. Furthermore, this computation would be highly inefficient since the *intrinsic* dimensionality or major degrees-of-freedom of $\tilde{\mathbf{U}}$ for each class is likely to be significantly smaller than N.

However, as derived in Chapter 3 using the subspace method, the complete high-dimensional likelihood estimate can be written as the product of two independent marginal Gaussian densities

$$\begin{aligned} \hat{P}(\tilde{\mathbf{U}}|\Omega) &= \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\epsilon^2(\tilde{\mathbf{U}})}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \\ &= P_F(\tilde{\mathbf{U}}|\Omega)\ \hat{P}_{\bar{F}}(\tilde{\mathbf{U}}|\Omega) \end{aligned} \tag{5.11}$$

where $P_F(\tilde{\mathbf{U}}|\Omega)$ is the true marginal density in F, $\hat{P}_{\bar{F}}(\tilde{\mathbf{U}}|\Omega)$ is the estimated marginal density in the orthogonal complement $\bar{F}$, y_i are the principal components and $\epsilon^2(\tilde{\mathbf{U}})$ is the residual (or DFFS).



(a) (b)

Figure 5-4: Examples of FERET frontal-view image pairs used for (a) the Gallery set (training) and (b) the Probe set (testing).

5.5 Experiments

To test our recognition strategy we used a collection of images from the FERET face database. This collection of images consists of hard recognition cases that have proven difficult for all face recognition algorithms previously tested on the FERET database. The difficulty posed by this dataset appears to stem from the fact that the images were taken at different times, at different locations, and under different imaging conditions. The set of images consists of pairs of frontal-views and are divided into two subsets: the "gallery" (training set) and the "probes" (testing set). The gallery images consisted of 74 pairs of images (2 per individual) and the probe set consisted of 38 pairs of images, corresponding to a subset of the gallery members. These images are shown in Figure 5-4.

Before we can apply our deformable matching technique, we need to perform a rigid alignment of these facial images. For this purpose we have used an automatic face-processing system which extracts faces from the input image and normalizes for translation, scale as well as slight rotations (both in-plane and out-of-plane). As described in Chapter 4 the system uses maximum-likelihood estimation of object location (in this case the position and scale of a face and the location of individual facial features) to geometrically align faces into standard normalized form as shown in Figure 5-5. All the faces in our experiments were geometrically aligned and normalized

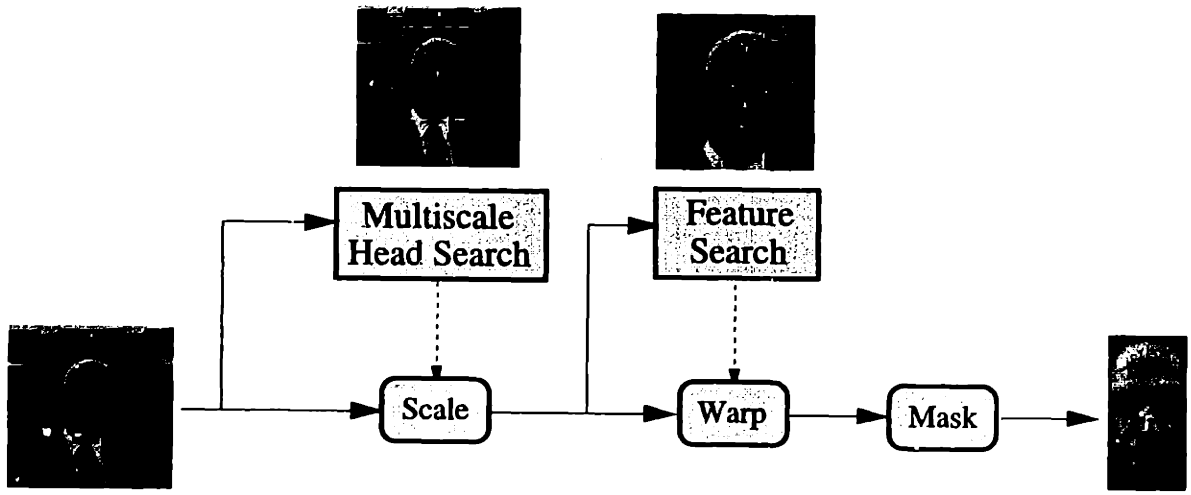


Figure 5-5: The face alignment system

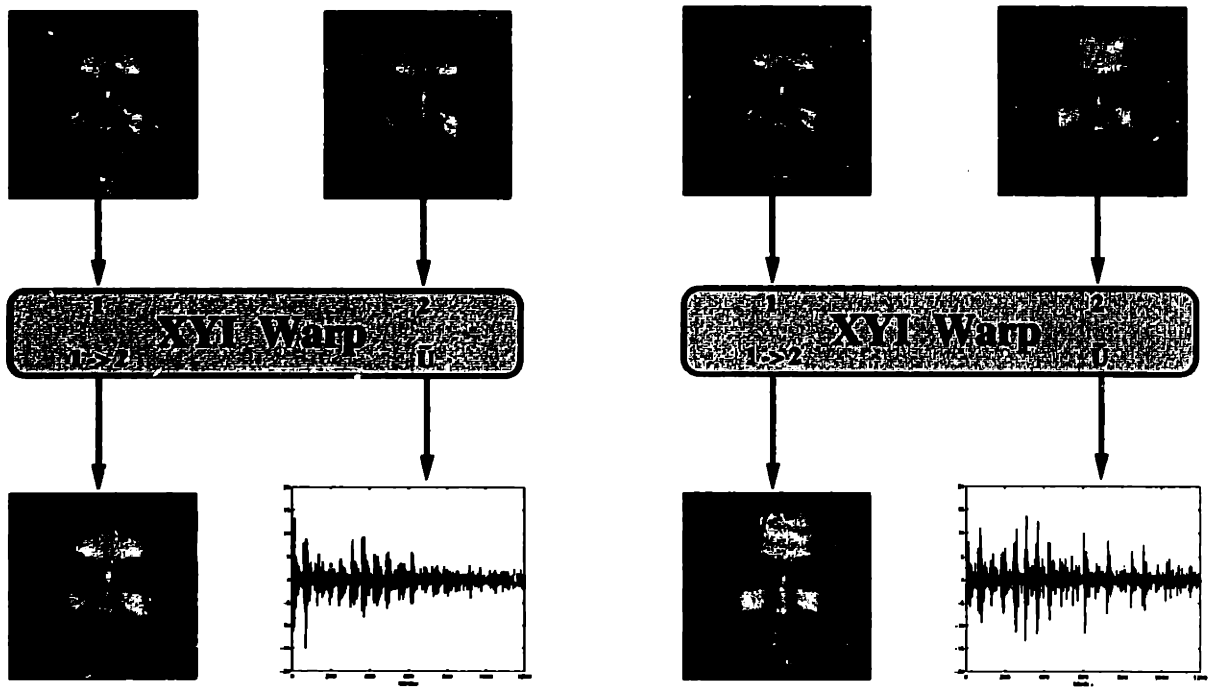
Figure 5-6: The first 8 normalized eigenfaces.

in this manner prior to further analysis.

5.5.1 Matching with Eigenfaces

As a baseline comparison, we first used an eigenface matching technique for recognition. The normalized images from the gallery and the probe sets were projected onto a 100-dimensional eigenspace and a nearest-neighbor rule based on a Euclidean distance measure was used to match each probe image to a gallery image.[1] A few of the lower-order eigenfaces used for this projection are shown in Figure 5-6. We note that these eigenfaces represent the principal components of an entirely different set of images — *i.e.*, none of the individuals in the gallery or probe sets were used in obtaining these eigenvectors. In other words, neither the gallery nor the probe sets were

[1] We note that this method corresponds to a generalized template-matching method which uses a Euclidean norm type of similarity $S(I_1, I_2)$, which is restricted to the principal component subspace of the data.



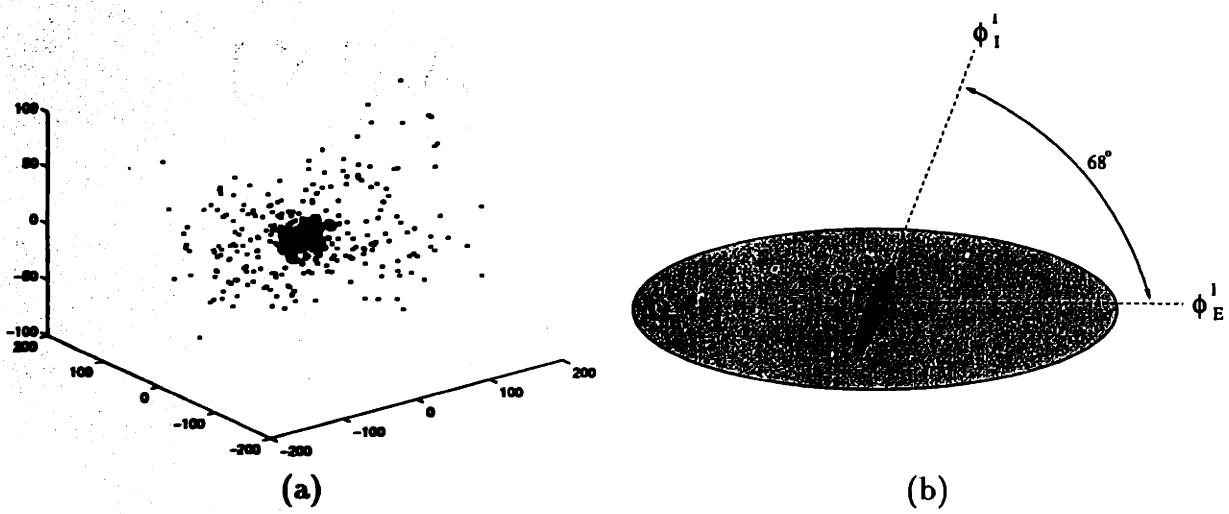
(a) (b)

Figure 5-7: Examples of (a) intrapersonal and (b) extrapersonal facial warps.

part of the "training set." The rank-1 recognition rate obtained with this method was found to be 84% (64 correct matches out of 76), and the correct match was always in the top 10 nearest neighbors. Note that this performance is better than or similar to recognition rates obtained by any algorithm tested on this database, and that it is lower (by about 10%) than the typical rates that we have obtained with the FERET database [24]. We attribute this lower performance to the fact that these images were selected to be particularly challenging. In fact, using an eigenface method to match the first views of the 76 individuals in the gallery to their second views, we obtain a higher recognition rate of 89% (68 out of 76), suggesting that the gallery images represent a less challenging data set since these images were taken at the same time and under identical lighting conditions.

5.5.2 Matching with XYI Deformations

For our probabilistic algorithm, we first gathered training data by computing the modal amplitude spectra for a training subset of 74 intrapersonal warps (by match-



(a) (b)

Figure 5-8: (a) distribution of the two classes in the first 3 principal components (circles for Ω_I, dots for Ω_E) and (b) schematic representation of the two distributions showing orientation difference between the corresponding principal eigenvectors.

ing the two views of every individual in the gallery) and a random subset of 296 extrapersonal warps (by matching images of *different* individuals in the gallery), corresponding to the classes Ω_I and Ω_E, respectively. An example of each of these two types of warps is shown in Figure 5-7.

It is interesting to consider how these two classes are distributed, for example, are they linearly separable or embedded distributions? One simple method of visualizing this is to plot their mutual principal components — *i.e.*, perform PCA on the *combined* dataset and project each vector onto the principal eigenvectors. Such a visualization is shown in Figure 5-8(a) which is a 3D scatter plot of the first 3 principal components. This plot shows what appears to be two completely enmeshed distributions, both having near-zero means and differing primarily in the amount of scatter, with Ω_I displaying smaller modal amplitudes as expected. It therefore appears that one can not reliably distinguish low-amplitude extrapersonal warps (of which there are many) from intrapersonal ones.

However, direct visual interpretation of Figure 5-8(a) is very misleading since we are essentially dealing with low-dimensional (or "flattened") hyper-ellipsoids which are intersecting near the origin of a very high-dimensional space. The key distinguishing factor between the two distributions is their relative orientation. Fortunately, we

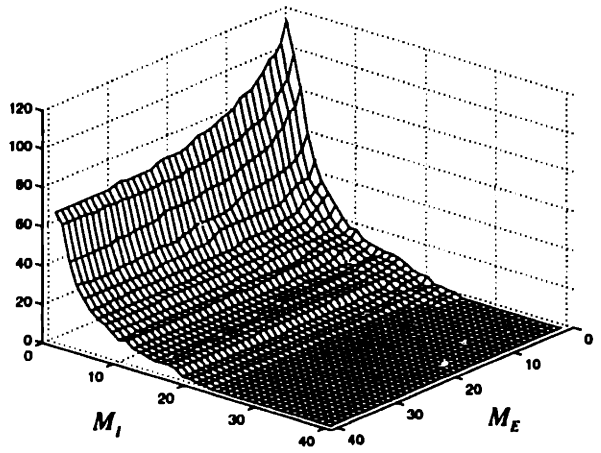


Figure 5-9: Total number of misclassified extrapersonal matches (with $P(\Omega_I|\tilde{\mathbf{U}}) > 0.5$) as a function of the principal subspace dimensionalities M_I and M_E.

can easily determine this relative orientation by performing a separate PCA on each class and computing the dot product of their respective first eigenvectors. This analysis yields the cosine of the angle between the major axes of the two hyper-ellipsoids, which was found to be 68°, implying that the orientation of the two hyper-ellipsoids is quite different. Figure 5-8(b) is a schematic illustration of the geometry of this configuration, where the hyper-ellipsoids have been drawn to approximate scale using the corresponding eigenvalues.

We note that since these classes are not linearly separable, simple linear discriminant techniques (*e.g.*, using hyperplanes) can not be used with any degree of reliability. The proper decision surface is inherently nonlinear (quadratic, in fact, under the Gaussian assumption) and is best defined in terms of the *a posteriori* probabilities — *i.e.*, by the equality $P(\Omega_I|\tilde{\mathbf{U}}) = P(\Omega_E|\tilde{\mathbf{U}})$. Fortunately, the optimal discriminant surface is automatically implemented when invoking a MAP classification rule.

Having analyzed the geometry of the two distributions, we then computed the likelihood estimates $P(\tilde{\mathbf{U}}|\Omega_I)$ and $P(\tilde{\mathbf{U}}|\Omega_E)$ using the PCA-based method outlined in Section 5.4.1. We selected principal subspace dimensions of $M_I = 10$ and $M_E = 30$ for Ω_I and Ω_E, respectively. These density estimates were then used with a default setting of equal priors, $P(\Omega_I) = P(\Omega_E)$, to evaluate the *a posteriori* intrapersonal probability $P(\Omega_I|\tilde{\mathbf{U}})$ for matching probe images to those in the gallery.

	XYI-warp	I-diff	XY-flow
Mean Correct Recognition Rate	86.8 %	85.9 %	82.3 %
Max Correct Recognition Rate	92.1 %	89.5 %	86.8 %
Mean Number of False Matches	10	14	1
Max Number of False Matches	115	155	53

Table 5.1: Performance of Bayesian classifier with three different data representations: full XYI-warp, intensity differences (I-diff) and optical flow (XY-flow). Results are mean/maximum values over nearly 2000 experimental trials with varying M_I and M_E.

In order to avoid an unnecessarily large number of XYI warps, we only matched a probe image to the top 10 gallery images retrieved by the eigenface method. This significantly reduces the computational cost of our system, since computing eigenface similarity scores is negligible compared to computing XYI warps (the former takes several milliseconds whereas the latter takes approximately 20 seconds on an HP 735 workstation).

Therefore, for each probe image we computed a set of 10 probe-to-gallery warps and re-sorted the matching order, this time using the *a posteriori* probability $P(\Omega_I|\tilde{\mathbf{U}})$ as the similarity measure. This probabilistic ranking yielded an improved rank-1 recognition rate of 92% (70 out of 76). Furthermore, out of the 608 extrapersonal warps performed in this recognition experiment, only 2% (11) were misclassified as being intrapersonal — *i.e.*, with $P(\Omega_I|\tilde{\mathbf{U}}) > P(\Omega_E|\tilde{\mathbf{U}})$.

We also analyzed the sensitivity of our Bayesian matching technique with respect to the principal subspace dimensionalities M_I and M_E, which are used in estimating the likelihoods $P(\tilde{\mathbf{U}}|\Omega_I)$ and $P(\tilde{\mathbf{U}}|\Omega_E)$. The higher we set these parameters the more accurate an estimate of the likelihoods we obtain, while also requiring more principal projections. These parameters therefore represent an accuracy *vs.* complexity tradeoff in our Bayesian approach. To quantify this tradeoff, we repeated the probe set recognition experiment while varying both parameters and noted that the recognition rate never dropped below 79%, even when the two subspaces used in estimating the likelihoods were as low as one-dimensional. However, we noted that the total num-

ber of extrapersonal matches which were misclassified as being intrapersonal — *i.e.*, $P(\Omega_I|\tilde{\mathbf{U}}) > P(\Omega_E|\tilde{\mathbf{U}})$ — varied in a principled way with the subspace dimensionalities. This variation is shown in Figure 5-9 and is clearly the type of behavior one would expect: the total number of misclassified matches decreases with increasing subspace dimensionalities. From the figure, it is apparent that these errors are more sensitively dependent on M_I, the dimensionality of the intrapersonal subspace (possibly because this class has a much lower *intrinsic* dimensionality and its distribution can be modeled using fewer principal eigenvectors).

5.5.3 Matching with Optical Flow and Intensity Differences

To compare the efficacy of our deformable representation for $d(I_1, I_2)$ (*i.e.*, the modal amplitudes of an XYI-warp), we next applied our Bayesian matching technique on the alternative representations: intensity differences and optical flow. The particular optical flow algorithm used in our experiment was that of Wang & Adelson [46]. For each method, the eigenspace analysis was used to derive corresponding density estimates for the intra/extra classes and recognition proceeded exactly as described in the previous section.

Since it is difficult to compare recognition and false match rates directly (due to the different dimensionalities of $d(I_1, I_2)$ in each case) we systematically varied the dimensions of the principal subspaces M_I and M_E, as in Figure 5-9 for each method and analyzed the performance in terms of % correct recognition and the number of false matches. Table 5.1 shows the mean and maximum values computed over the nearly 2,000 different combinations of M_I and M_E for the three different methods: full XYI-warp, intensity differences (I-diff) and optical flow (XY-flow). These results indicate that XYI-warps are in fact the best representation for classification purposes, with intensity differences being second and optical flow being the least effective representation. We believe the reason optical flow is so ineffective is because it has no intensity information encoded in the representation and also since it essentially yields "garbage" for the extrapersonal class (due to the inability of obtaining good correspondences between two different individuals). Notice how the number of

false matches, however, is least with optical flow, possibly because it is quite easy to discriminate between the (essentially "garbage") flow field of an extrapersonal warp and that of an intrapersonal one. Also note that in terms of false matches, intensity differences seem to yield worse results than XYI-warps.

Chapter 6

FERET Test Results

In this chapter we present results of the face recognition system on experimental trials conducted as part of the FERET program.

6.1 The FERET Program

The Face Recognition Technology (FERET) program is sponsored by the US Department of Defense's Counterdrug Technology Transfer Program with the US Army Research Laboratory (ARL) serving as the technical agent. The FERET database and testing procedure is designed to assess the state of the art in face recognition by providing a standardized testing platform for researchers, such that recognition rates can be objectively compared. The database is divided into a development set given to the researchers and a sequestered test set used for evaluation.

The images in the database were acquired from a 35mm camera and processed for digital storage and distribution. Each image was labeled with respect to the identity and pose of the individuals. Images of individuals were acquired in several different formats. A set of frontal images labeled FA and FB, corresponding to two views taken moments apart with different expressions, and so called "duplicate" images which are views of the same individual taken at different times (weeks, months and upto a year apart). In addition to these frontal views, the database also contains non-frontal poses including profiles. As of July 1996, a total of over 14,000 images had been captured

corresponding to 1,200 individuals. From this data, a subset of approximately 500 sets of images were distributed to researchers for purposes of training.

The standard testing procedure consists of matching images in two separate sets, a gallery set and a probe set. For every image in the probe set, a similarity score is computed to every entry in the gallery set. These similarity scores are then sorted and ranked and used to compute a "cumulative match score" as shown for example in Figure 6-2 which shows the number of correct matches in the top N most similar entries computed by the recognition algorithm. The typical recognition rates reported in the literature therefore correspond to the rank-1 match or the first point on this graph.

The first FERET test took place in August 1994 and established a baseline for face recognition algorithms consisting of both frontal FA/FB as well as non-frontal views. This was followed in March 1995 by a larger gallery test which emphasized duplicates. Both these tests required automatic face detection in addition to recognition. These tests were followed by the September 1996 test which no longer required automatic detection, providing ground truth locations for the eyes. This test consisted of matching over 3000 frontal images containing both frontal FA/FBs as well as duplicates.

6.2 MIT Algorithm Performance

There were essentially two different versions of the MIT face recognition system which were tested during a three year span. The first consisted of a standard eigenface nearest neighbor recognition technique which was used in the 1994/1995 tests. The second, improved system differed by use of the Bayesian similarity measure in place of the nearest neighbor matching rule and was used in the 1996 test.

For computational reasons, the representation selected for $d(I_1, I_2)$ in the 1996 version was that of intensity differences, *i.e.* $I_1 - I_2$ rather than the full XYI warping method. We note that the two mutually exclusive classes Ω_I and Ω_E corresponding to the intrapersonal and extrapersonal image differences result in the dual eigenfaces

Institution	Recognition Rate
MIT Media Lab (August 1996)	96
Rockefeller (November 1995)	96
USC (March 1995)	92
MIT Media Lab (March 1995)	88

Table 6.1: FA vs FB results on the FERET 1995/1996 tests

Institution	Recognition Rate
MIT Media Lab (August 1996)	69
Rockefeller (November 1995)	62
USC (March 1995)	58
MIT Media Lab (March 1995)	40

Table 6.2: Duplicate Scores on the FERET 1995/1996 tests

shown in Figure 6-1. Note that the intrapersonal variations shown in Figure 6-1-(a) represents subtle variations due mostly to expression changes whereas the extrapersonal variations in Figure 6-1-(b) are more representative of general eigenfaces which code variations such as hair color, facial hair and glasses.

Tables 6.1 and 6.2 show a comparison of the rank-1 recognition accuracy of the algorithms that competed in the 1995 FERET test on frontal FAs and FBs as well as the duplicates. The MIT 1995 system is the nearest neighbor matching technique and the 1996 version is the Bayesian similarity technique using the dual intra-extra eigenfaces with intensity differences. Note the increased performance obtained using the Bayesian similarity measure especially in regards to the harder problem of duplicate images.

The performance contrast between the nearest neighbor and the Bayesian similarity techniques is further illustrated in Figures 6-2 and 6-3 which show the cumulative match scores. Note an approximate 8% improvement in the case of FA/FB and a dramatic 30% improvement for duplicate images.

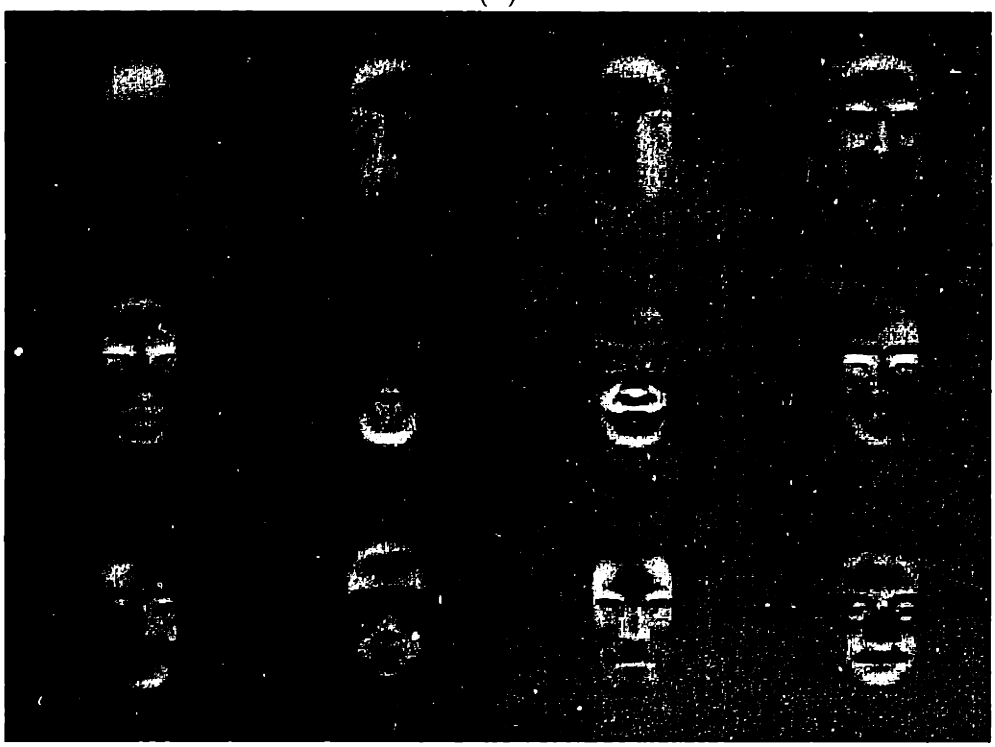
In September 1996, both versions of our algorithm were tested along with four other competitors on a large gallery test consisting of over 3000 images. Figures 6-

Algorithm	gallery 1	gallery 2	gallery 3	gallery 4	gallery 5
ARL Eigenface	6	6	3	2	5
ARL Correlation	7	4	4	4	6
Excalibur Corp.	2	3	2	3	1
MIT Sep 96	1	1	1	1	1
MIT Mar 95	4	2	5	6	7
Rutgers Univ.	3	4	7	5	4
Univ. of Maryland	4	6	6	7	1
Average	0.220	0.587	0.626	0.512	0.653
Number of Probes Scored	143	64	194	277	44

Table 6.3: Variations in performance over 5 different galleries of fixed size(200) on duplicate probes. Algorithms are order by performance (1 to 7). The order is by percentage of probes correctly identified (rank 1). Also included in the table is average rank 1 performance for all algorithms and number of probes scored

4 and 6-5 show a comparison of the various algorithms on FA/FB and duplicate images. Note that the MIT96 algorithm outperforms all the other competitors by a margin of approximately 10%. Table 6.3 shows the overall ranking of the competitors on different subsets of the test. The data in this table and other figures were taken from the ARL report on the FERET test results [37].

(a)



(b)

Figure 6-1: Dual Eigenfaces: (a) Intrapersonal, (b) Extrapersonal

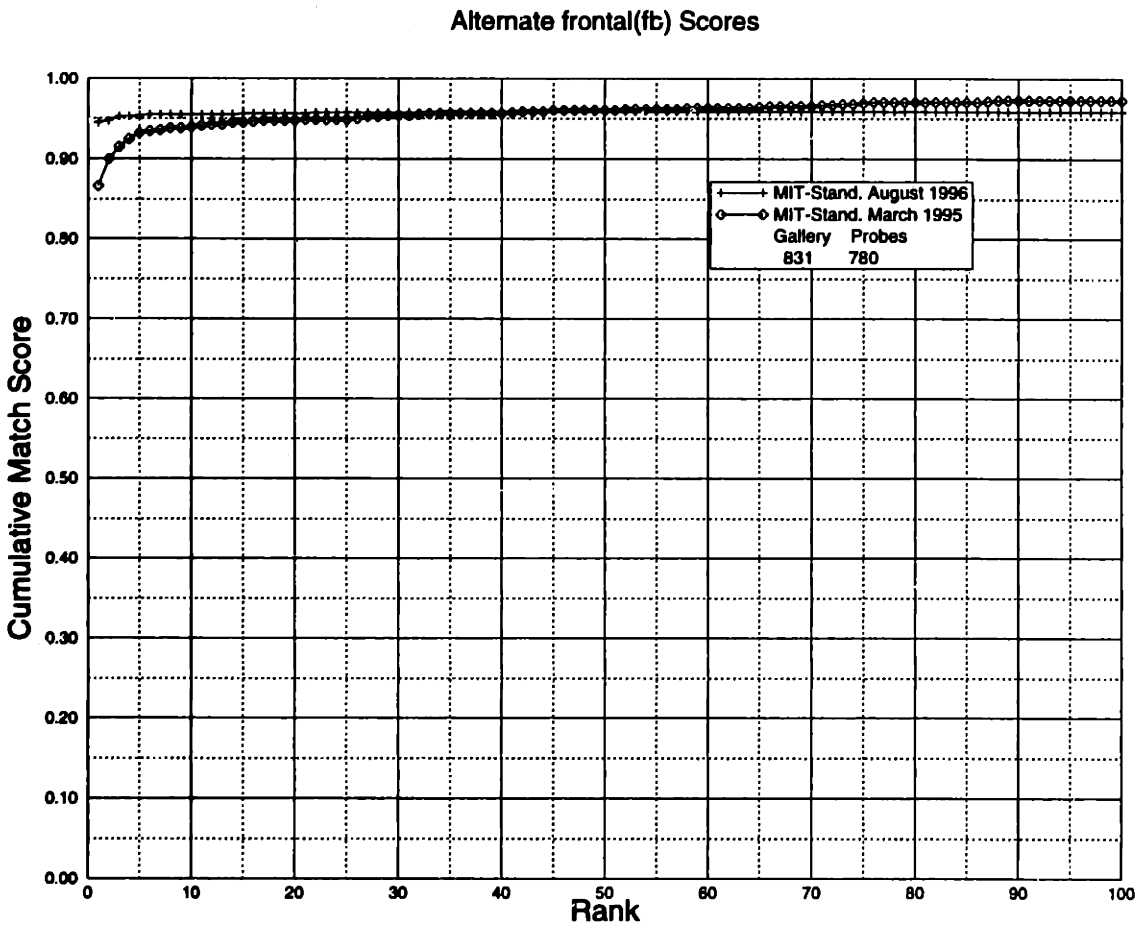


Figure 6-2: Comparison of nearest-neighbor (MIT95) vs. Bayesian similarity (MIT96) methods on FA/FB FERET data.

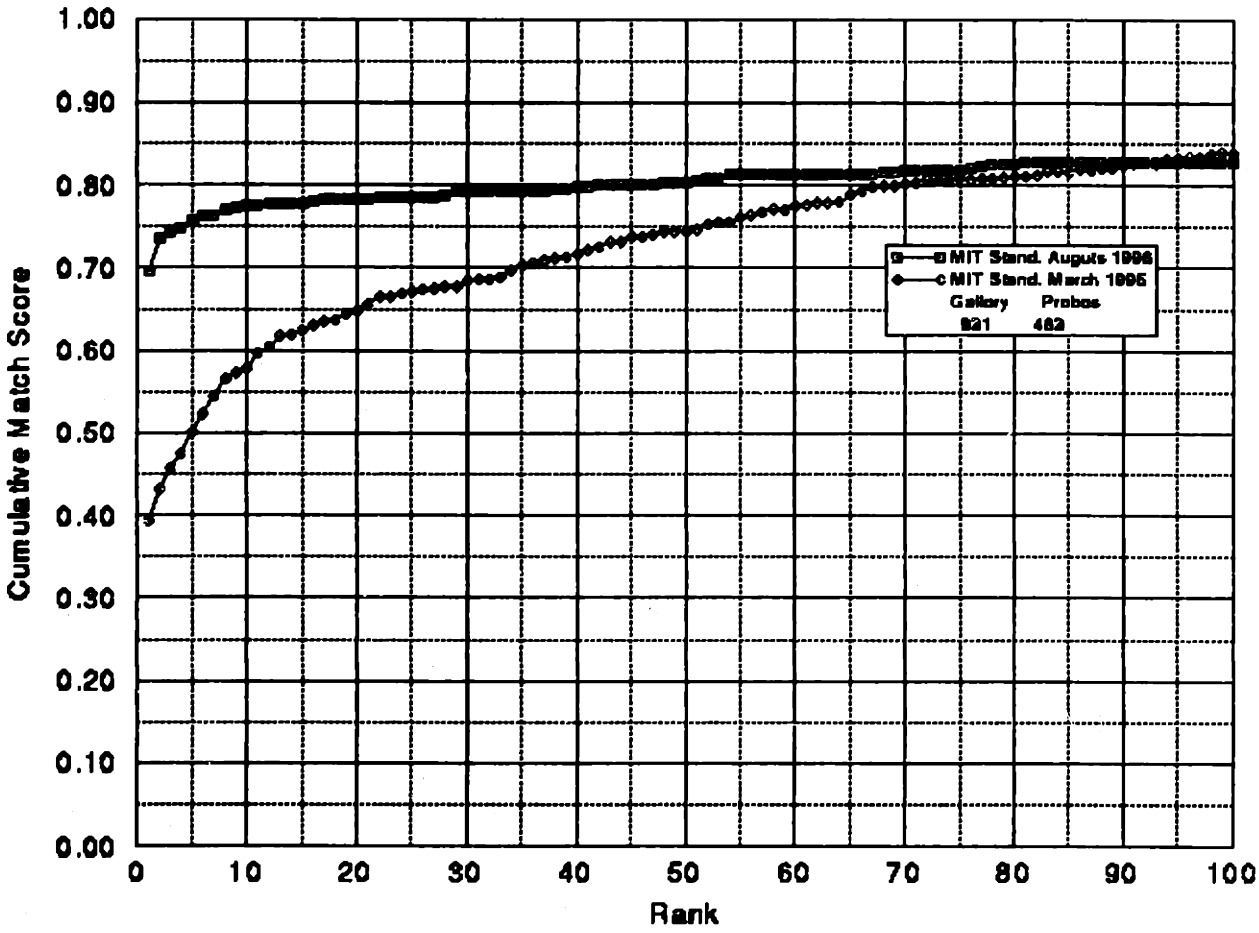


Figure 6-3: Comparison of nearest-neighbor (MIT95) vs. Bayesian similarity (MIT96) methods on Duplicate FERET data.

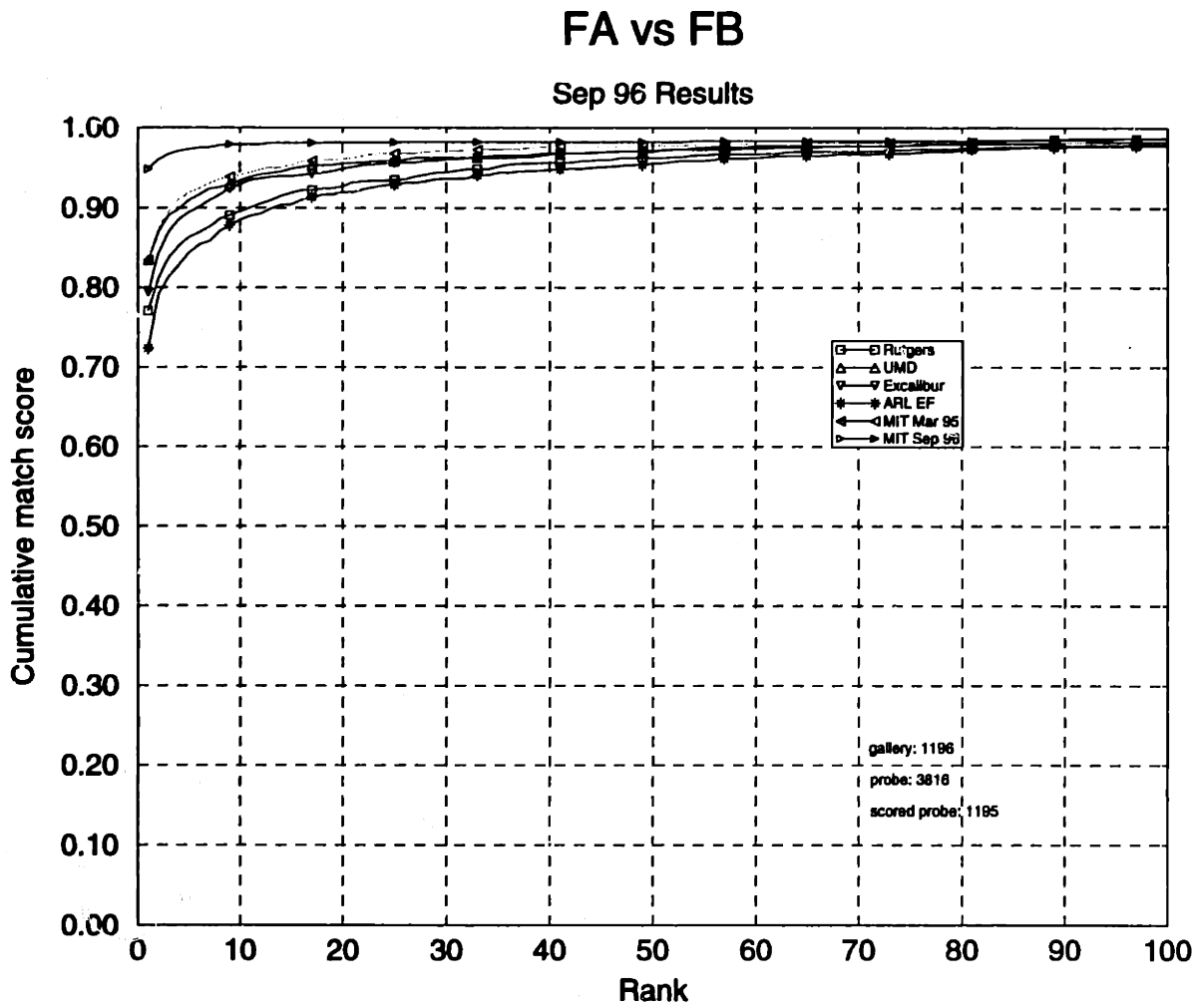


Figure 6-4: Results of FERET'96 Competition on FA/FB data.

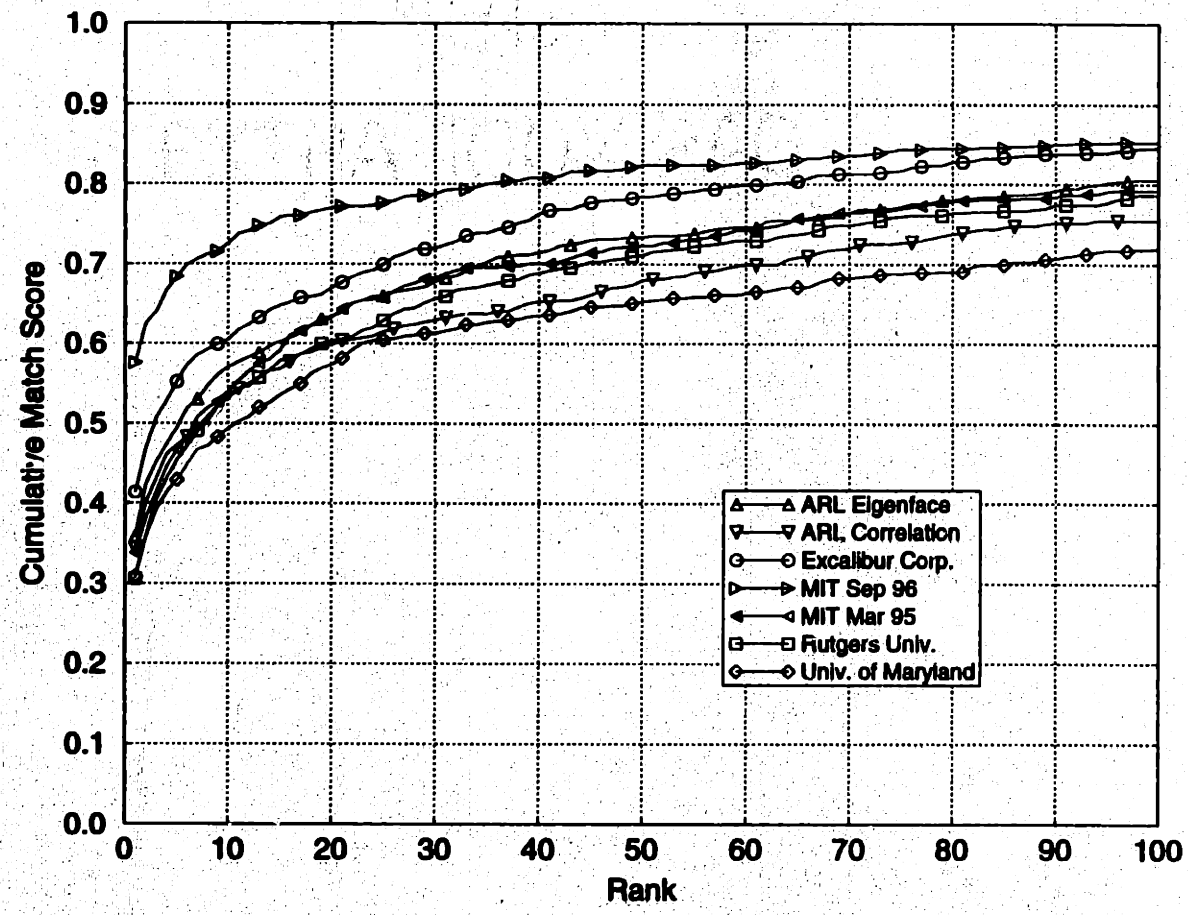


Figure 6-5: Results of FERET'96 Competition on Duplicate data.

Chapter 7

Conclusions & Future Work

7.1 Estimation & Detection

We have described a density estimation technique for unsupervised visual learning which exploits the *intrinsic* low-dimensionality of the training imagery to form a computationally simple estimator for the complete likelihood function of the object. Our estimator is based on a subspace decomposition and can be evaluated using only the M-dimensional principal component vector. We have derived the form for an optimal estimator and its associated expected cost for the case of a Gaussian density. In contrast to previous work on learning and characterization — which uses PCA primarily for dimensionality reduction and/or feature extraction — our method uses the eigenspace decomposition as an integral part of estimating *complete* density functions in high-dimensional image spaces. These density estimates were then used in a maximum likelihood formulation for target detection. The multiscale version of this detection strategy was demonstrated in applications in which it functioned as an attentional subsystem for object recognition. The performance was found to be superior to existing detection techniques in experimental results on a large number of test data (on the order of thousands).

7.2 Recognition

We have also proposed an alternative technique for direct visual matching of images for purposes of recognition and database search. Specifically, we have argued in favor of a *probabilistic* measure of similarity, in contrast to simpler methods which are based on standard L_2 norms (*e.g.*, template matching) or subspace-restricted norms (*e.g.*, eigenspace matching). This probabilistic framework is also advantageous in that the intra/extra density estimates explicitly characterize the type of appearance variations which are critical in formulating a meaningful measure of similarity. For example, the deformations corresponding to facial expression changes (which may have high image-difference norms) are, in fact, *irrelevant* when the measure of similarity is to be based on *identity*. The subspace density estimation method used for representing these classes thus corresponds to a *learning* method for discovering the principal modes of variation important to the classification task. Furthermore, by equating similarity with the *a posteriori* probability $P(\Omega_I|d(I_1, I_2))$, we obtain an optimal non-linear decision rule for matching and recognition. This aspect of our approach differs from methods which use linear discriminant analysis techniques for visual object recognition (*e.g.*, [47]).

Furthermore, in Chapter 5 we have experimentally shown that our deformable XYI warping method for obtaining pixel correspondences does indeed lead to an effective representation for $d(I_1, I_2)$, especially when compared with simpler methods such as intensity differences and optical flow. In fact, these methods can essentially be viewed as limiting cases of our general XYI warping method and therefore lack full correspondence: the intensity difference method requires pre-established spatial correspondence between I_1 and I_2, whereas optical flow assumes that I_1 and I_2 only differ by an XY deformation. The XYI warping method, on the other hand, makes no such assumptions and efficiently solves for both types of correspondences in a unified framework. The resultant modal amplitude spectra of these deformations will therefore encode both shape (spatial) and texture (intensity) variations between the two images. The experimental results indicate that a $d(I_1, I_2)$ representation based

on full XYI correspondence (*i.e.*, precise alignment/correspondence) does in fact lead to the best overall recognition performance.

Nevertheless, the 1996 FERET test results shown in Chapter 6 have shown that, given the rough initial alignment obtained with the face processor, a $d(I_1, I_2)$ representation based on intensity differences is in fact sufficient for achieving high recognition rates while requiring a considerably lesser amount of computation than a full XYI warp.

7.3 Future Work

We note that from a probabilistic perspective, the class-conditional density $P(\mathbf{x}|\Omega)$ is the most important data representation to be learned. This density is the critical component in detection, recognition, prediction, interpolation and general inference. For example, having learned these densities for several object classes $\{\Omega_1, \Omega_2, \cdots, \Omega_n\}$, one can invoke a Bayesian framework for classification and recognition:

$$P(\Omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega_i)P(\Omega_i)}{\sum_{j=1}^{n} P(\mathbf{x}|\Omega_j)P(\Omega_j)} \tag{7.1}$$

Such a framework is also important in detection. In fact, the ML detection framework can be extended using the notion of a "not-class" $\bar{\Omega}$, resulting in *a posteriori* saliency maps of the form

$$P(\Omega|\mathbf{x}) = \frac{P(\mathbf{x}|\Omega)P(\Omega)}{P(\mathbf{x}|\bar{\Omega})P(\bar{\Omega}) + P(\mathbf{x}|\Omega)P(\Omega)} \tag{7.2}$$

where now a maximum *a posteriori* (MAP) rule can be used to estimate the position and scale of the object. One difficulty with such a formulation is that the "not-class" $\bar{\Omega}$ is, in practice, too broad a category and is therefore multimodal and very high-dimensional. One possible approach to this problem is to use ML detection to identify the particular subclass of $\bar{\Omega}$ which has high likelihoods (*e.g.*, typical false alarms) and then to estimate this distribution and use it in the MAP framework. This

can be viewed as a probabilistic approach to learning using positive as well as *negative* examples. The use of negative examples has been shown to be critically important in building robust face detection systems by Sung and Poggio [43]. Similarily, the face finding system of Rowley [41] uses a neural network classifier as opposed to a Bayesian discriminator. It should be interesting to compare the detection performance of the neural network technique to that of a probabilistic one, since in the limit of infinite training data, a neural network should equal the performance of a Bayesian classifier.

Another possible extension of the density estimation methods in this thesis would be modeling temporal data such as hand gestures and body movements. With suitable time normalization (through dynamic time warping techniques) it should be possible to apply the same detection and recognition methods used with 2D imagery to multidimensional time series. A similar extension would be to volumetric data such as medical images.

Appendix A

The cost function we wish to minimize is the KL divergence

$$D(p|\hat{p}) = \int p(x) \log \left(p(x)/\hat{p}(x) \right) d(x) \tag{A.1}$$

which is simply the expectation of the log ratio and given the two Gaussian densities in Equation 3.11 reduces to

$$\begin{aligned}
\log \frac{p(x)}{\hat{p}(x)} &= -\frac{d(x)}{2} - \frac{N}{2} \log (2\pi) - \frac{1}{2} \log |\Sigma| + \frac{d'(x)}{2} + \frac{N}{2} \log (2\pi) + \frac{1}{2} \log |\Sigma'| \\
&= \frac{1}{2} \left[d'(x) - d(x) - \log |\Sigma| + \log |\Sigma'| \right] \\
&= \frac{1}{2} \left[\sum_{i=1}^{M} \frac{\mathbf{y}_i^2}{\lambda_i} + \frac{1}{\rho} \sum_{i=M+1}^{N} \mathbf{y}_i^2 - \sum_{i=1}^{M} \frac{\mathbf{y}_i^2}{\lambda_i} - \sum_{i=M+1}^{N} \frac{\mathbf{y}_i^2}{\lambda_i} + \log |\Sigma'| - \log |\Sigma| \right] \\
&= \frac{1}{2} \left[\frac{1}{\rho} \sum_{i=M+1}^{N} \mathbf{y}_i^2 - \sum_{i=M+1}^{N} \frac{\mathbf{y}_i^2}{\lambda_i} + \left(\sum_{i=1}^{M} \log \lambda_i + \sum_{i=M+1}^{N} \log \rho \right) - \left(\sum_{i=1}^{M} \log \lambda_i + \sum_{i=M+1}^{N} \log \lambda_i \right) \right] \\
&= \frac{1}{2} \left[\frac{1}{\rho} \sum_{i=m+1}^{N} \mathbf{y}_i^2 - \sum_{i=M+1}^{N} \frac{\mathbf{y}_i^2}{\lambda_i} + \sum_{i=M+1}^{N} \log \rho - \sum_{i=M+1}^{N} \log \lambda_i \right] \\
&= \frac{1}{2} \sum_{i=M+1}^{N} \left[\frac{\mathbf{y}_i^2}{\rho} - \frac{\mathbf{y}_i^2}{\lambda_i} + \log \rho - \log \lambda_i \right]
\end{aligned} \tag{A.2}$$

Therefore

$$J(\rho) = D(p|\hat{p}) = E \left[\log \frac{p(x)}{\hat{p}(x)} \right] = \frac{1}{2} \sum_{i=M+1}^{N} \left[\frac{\lambda_i}{\rho} - 1 + \log \frac{\rho}{\lambda_i} \right] \tag{A.3}$$

Taking the derivative and setting it equal to zero we obtain

$$\begin{aligned}\frac{\delta D}{\delta \rho} &= \frac{1}{2}\sum_{i=M+1}^{N}\frac{\delta}{\delta\rho}\left[\frac{\lambda_i}{\rho}-1+\log\frac{\rho}{\lambda_i}\right] \\ &= \frac{1}{2}\sum_{i=M+1}^{N}\left[-\frac{\lambda_i}{\rho^2}+\frac{1}{\rho}\right] \\ &= 0\end{aligned} \tag{A.4}$$

which implies

$$\rho^* = \frac{1}{N-M}\sum\lambda_i \tag{A.5}$$

To show that this is a minimum we check the 2nd derivative

$$\begin{aligned}\frac{\delta^2 D}{\delta\rho^2} &= \frac{1}{2}\sum_i\frac{\delta}{\delta\rho}\left[\frac{\lambda_i}{\rho}-1+\log\frac{\rho}{\lambda_i}\right] \\ &= \frac{1}{2}\sum_i\left[\frac{\lambda_i}{\rho^3}-\frac{1}{\rho^2}\right] \\ &= \frac{1}{2\rho^2}\sum_i\left[\frac{2\lambda_i}{\rho}-1\right] \\ &= \frac{1}{2\rho^2}\left[\frac{2}{\rho}\sum\lambda_i-(N-M)\right] \\ &= \frac{N-M}{2\rho^{*2}}\left[\frac{2}{\rho^*}\frac{\sum\lambda_i}{N-M}-1\right] \quad \text{since } \rho=\rho^*=\tfrac{1}{N-M}\textstyle\sum\lambda_i \\ &= \frac{N-M}{2\rho^{*2}}\left[\frac{2}{\rho^*}\rho^*-1\right] \\ &= \frac{N-M}{2\rho^{*2}} \\ &> 0\end{aligned} \tag{A.6}$$

Next we show the unbiasedness of the estimator based of ρ^* with a simple derivation.

$$d = \sum_{i=1}^{N}\frac{\mathbf{y}_i^2}{\lambda_i} \tag{A.7}$$

has expectation N

$$\begin{aligned}E[d] &= \sum_{i=1}^{N}\frac{E[\mathbf{y}_i^2]}{\lambda_i} \\ &= \sum_{i=1}^{N}\frac{\lambda_i}{\lambda_i} \\ &= N\end{aligned} \tag{A.8}$$

An estimator based on a single eigenvalue ρ has the form

$$\hat{d} = \frac{1}{\rho}\sum_{i=1}^{N} \mathbf{y}_i^2 \tag{A.9}$$

with expectation

$$\begin{aligned} E[\hat{d}] &= \frac{1}{\rho}\sum_{i=1}^{N} E[\lambda_i] \\ &= \frac{1}{\rho}\sum_{i=1}^{N} \lambda_i \end{aligned} \tag{A.10}$$

If we then require unbiasedness, $E[d] = E[\hat{d}]$, then

$$\frac{1}{\rho}\sum_{i=1}^{N} \lambda_i = N \tag{A.11}$$

which yields

$$\rho = \frac{1}{N}\sum_{i=1}^{N} \lambda_i \tag{A.12}$$

Thus a value of ρ based on the average of eigenvalues of a subspace will yield an unbiased estimate.

Bibliography

[1] C.H. Anderson, P.J. Burt, and G.S. Van der Wall. Change detection and tracking using pyramid transform techniques. In *Proc. of SPIE Conf. on Intelligence, Robots and Computer Vision*, volume 579, pages 72–78, 1985.

[2] K. J. Bathe. *Finite Element Procedures in Engineering Analysis.* Prentice-Hall, 1982.

[3] David Beymer. Vectorizing face images by interleaving shape and texture computations. A.I. Memo No. 1537, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1995.

[4] M.J. Black and A.D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using view-based representation. In *European Conference on Computer Vision*, Cambridge, England, April 1996.

[5] C. Bregler and S.M. Omohundro. Surface learning with applications to lip reading. In G. Tesauro J.D. Cowan and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 43–50. Morgan Kaufman Publishers, San Fransisco, 1994.

[6] R. Brunelli and S. Messelodi. Robust estimation of correlation: An application to computer vision. Technical Report 9310-015, IRST, October 1993.

[7] M.C. Burl, U.M. Fayyad, P. Perona, P. Smyth, and M.P. Burl. Automating the hunt for volcanos on venus. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, Seattle, WA, June 1994.

[8] T.F. Cootes and C.J. Taylor. Active shape models: Smart snakes. In *Proc. British Machine Vision Conference*, pages 9–18. Springer-Verlag, 1992.

[9] M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1994.

[10] I. Craw and P. Cameron. Face recognition by computer. In D. Hogg and R. Boyle, editors, *Proc. British Machine Vision Conference*, pages 498–507. Springer-Verlag, 1992.

[11] I. Craw and et al. Automatic face recognition: Combining configuration and texture. In Martin Bichsel, editor, *Proc. Int'l Workshop on Automatic Face and Gesture Recognition*, Zurich, 1995.

[12] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1977.

[13] C.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins Press, 1989.

[14] B.K.P. Horn and G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[15] M.K. Hu. Visual pattern recognition by moment invariants. *IEEE Trans. on Information Theory*, 8:179–187, 1962.

[16] I.T. Jolliffe. *Principal Component Analysis.* Springer-Verlag, New York, 1986.

[17] M. J. Jones and T. Poggio. Model-based matching by linear combination of prototypes. AI Memo No. 1583, Artificial Intelligence Laboratory, Massachusettes Institute of Technology, November 1996.

[18] T. Kanade. Picture processing by computer complex and recognition of human faces. Technical report, Kyoto University, Dept. of Information Science, 1973.

[19] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *Int'l Journal of Computer Vision*, 1(4):321–331, 1987.

[20] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1), 1990.

[21] B. Kumar, D. Casasent, and H. Murakami. Principal component imagery for statistical pattern recognition correlators. *Optical Engineering*, 21(1), 1982.

[22] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *IEEE Proceedings of the Fifth International Conference on Computer Vision (ICCV'95)*, Cambridge, MA, June 1995.

[23] M.M. Loeve. *Probability Theory.* Van Nostrand, Princeton, 1955.

[24] B. Moghaddam and A. Pentland. Face recognition using view-based and modular eigenspaces. *Automatic Systems for the Identification and Inspection of Humans*, 2277, 1994.

[25] H. Murase and S.K. Nayar. Visual learning and recognition of 3D objects from appearance. *Int'l Journal of Computer Vision*, 14(1), 1995.

[26] C. Nastar. Vibration modes for nonrigid motion analysis in 3D images. In *Proceedings of the Third European Conference on Computer Vision (ECCV '94)*, Stockholm, May 1994.

[27] C. Nastar and N. Ayache. Fast segmentation, tracking, and analysis of deformable objects. In *IEEE Proceedings of the Third International Conference on Computer Vision (ICCV'93)*, Berlin, May 1993.

[28] C. Nastar and A. Pentland. Matching and recognition using deformable intensity surfaces. In *IEEE International Symposium on Computer Vision*, Coral Gables, USA, November 1995.

[29] S.K. Nayar, S. Baker, and H. Murase. Parametric feature detection. In *Proc. of IEEE Conf. on Computer Vision & Pattern Recognition*, pages 471–477, San Francisco, CA, June 1996.

[30] S.K. Nayar, H. Murase, and S.A. Nene. General learning algorithm for robot vision. *Neural & Stochastic Methods in Image & Signal Processing*, 2304, 1994.

[31] S.E. Palmer. *The Psychology of Perceptual Organization: A Transformational Approach.* Academic Press, 1983.

[32] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. of IEEE Conf. on Computer Vision & Pattern Recognition*, Seattle, WA, June 1994.

[33] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. *Storage and Retrieval of Image and Video Databases II*, 2185, 1994.

[34] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recovery. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(7):715–729, 1991.

[35] A. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modelling and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13(7):715–729, July 1991.

[36] E. Persoon and K. S. Fu. Shape discrimination using fourier descriptors. *Proc. 2nd IJCPR*, pages 126–130, 1974.

[37] P. J. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proceedings Co. puter Vision and Pattern Recognition 97*, 1997.

[38] T. Poggio and F. Girosi. Networks for approximation and learning. In *Proceedings of the IEEE*, volume 78, pages 1481–1497, 1990.

[39] R.A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, 1984.

[40] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Detection of interest points using symmetry. In *Proc. of Int'l Conf. on Computer Vision*, Osaka, Japan, 1990.

[41] H. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical report, CMU-CS-95-158, Carnegie Mellon University, 1995.

[42] S. Sclaroff and A. Pentland. Modal matching for correspondence and recognition. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 17(6):545–561, 1995.

[43] K.K. Sung and T. Poggio. Example-based learning for view-based human face detection. In *Proc. of Image Understanding Workshop*, Monterey, CA, November 1994.

[44] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[45] J. M. Vincent, J. B. Waite, and D. J Myers. Automatic location of visual features by a system of multilayered perceptrons. In *IEE Proceedings*, volume 139, 1992.

[46] J.Y.A. Wang and E.H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, September 1994.

[47] J.J. Weng. On comprehensive visual learning. In *Proc. NSF/ARPA Workshop on Performance vs. Methodology in Computer Vision*, Seattle, WA, June 1994.