# Principal Manifolds and Probabilistic Subspaces for Visual Recognition

Baback Moghaddam, *Member*, *IEEE*

**Abstract**—We investigate the use of linear and nonlinear principal manifolds for learning low-dimensional representations for visual recognition. Several leading techniques: Principal Component Analysis (PCA), Independent Component Analysis (ICA), and nonlinear Kernel PCA (KPCA) are examined and tested in a visual recognition experiment using 1,800+ facial images from the "FERET" database. We compare the recognition performance of nearest-neighbor matching with each principal manifold representation to that of a maximum a posteriori (MAP) matching rule using a Bayesian similarity measure derived from dual probabilistic subspaces. The experimental results demonstrate the simplicity, computational economy, and performance superiority of the Bayesian subspace method over principal manifold techniques for visual matching.

**Index Terms**—Subspace techniques, PCA, ICA, Kernel PCA, Probabilistic PCA, learning, density estimation, face recognition.

## 1 Introduction

In recent years, computer vision research has witnessed a growing interest in subspace analysis techniques. In particular, eigenvector decomposition has been shown to be a highly effective tool for problems which have high-dimensional signal formats (e.g., an image array) but, nevertheless, represent visual phenomena which are intrinsically low-dimensional—needless to say, this is all too often the case in computer vision. Subspace analysis is heavily used in appearance-based modeling and recognition where the *principal modes* or the characteristic *degrees-of-freedom* are extracted and used for description, detection, and recognition. The identification and parametric representation of data in terms of these "principal manifolds" is to be found in physically-based modeling [37], correspondence and matching [41], parametric descriptions of shape [10], target detection [36], [7], [34], nonlinear image interpolation [5], appearance-based visual learning [32], [33], [35], [29], [30], face recognition [44], [36], [28], Linear Discriminant Analysis [13], Fisherfaces [3], as well as subspace density estimation [29], [30].

Subspace methods are often critical in machine learning where they are used to extract low-dimensional manifolds comprised of statistically uncorrelated or independent variables which tend to simplify tasks such as regression, classification, and density estimation. The Karhunen-Loève Transform (KLT) [22] and Principal Components Analysis (PCA) [18] are examples of eigenvector-based techniques which are commonly used for dimensionality reduction and feature extraction. Independent Factor Analysis (IFA) [1] and more specifically Independent Component Analysis (ICA) [9] is another linear decomposition which seeks statistically *independent* and non-Gaussian components, modeling the observed data as a linear mixture of (unknown) independent sources. ICA's proficiency in "blind source separation" [20] has found a particular niche in the analysis of EEG [23] and fMRI [25] signals of the brain. Nonlinear PCA (NLPCA) [21], [12], nonlinear Principal Surfaces [15], [16], "kernel" PCA [40], and nonlinear latent variable models [14] are various extensions of these linear techniques. In the following section, we will review some of these principal manifolds, their derivation, and consequent statistical properties. In Section 3, an alternative technique using probabilistic subspaces is presented and its performance is compared to principal manifolds in Section 4. We conclude with a discussion of the pros and cons of the different recognition techniques in Section 5.

## 2 Subspace Representations

Spatiotopic visual data (e.g., images, depth maps, flow fields, etc.) can be represented (after proper normalization) as vectors—i.e., as points in a high-dimensional vector space. For example, a $m$-by-$n$ 2D image $P(i,j)$ can be mapped to a vector $\mathbf{x} \in \mathcal{R}^{N=mn}$, by lexicographic ordering of the pixel elements.[1] Despite this high-dimensional embedding, the natural constraints of the physical world (and the imaging process) dictate that the data will, in fact, lie in a lower-dimensional (though possibly disjoint) manifold. The primary goal of subspace analysis is to identify, represent, and parameterize this manifold in accordance with some optimality criteria. In the next section, we review several techniques for computing both linear and nonlinear principal manifolds and highlight their corresponding statistical properties. It should be pointed out that, in this paper, we are assuming that the data can be modeled by a compact and connected (nondisjoint) manifold—which is often the case for frontal faces. More sophisticated techniques for subspace modeling of disjoint manifolds exist [30], [43], [4] which would, for example, be required if variations in pose and lighting were to modeled.

• *The author is with Mitsubishi Electric Research Laboratory, 201 Broadway, Cambridge, MA 02139. E-mail: baback@merl.com.*

*Manuscript received 16 Aug. 2000; revised 28 June 2001; accepted 15 Oct. 2001.*
*Recommended for acceptance by D. Kriegman.*
*For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112725.*

1. Without loss of generality, we will hereafter assume that the mean image vector $\bar{\mathbf{x}}$ is always subtracted from the data.

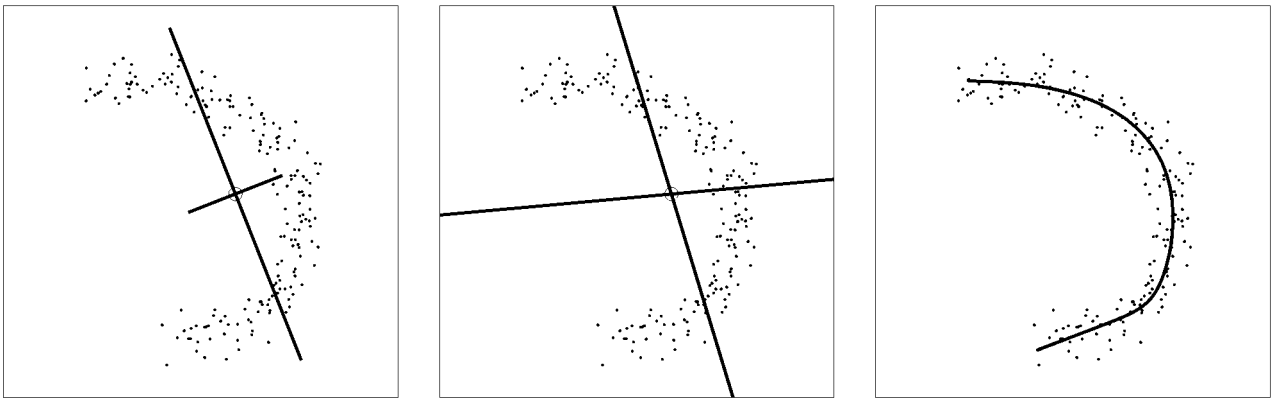



Fig. 1. (a) PCA basis (linear, ordered, and orthogonal). (b) ICA basis (linear, unordered, and nonorthogonal). (c) Principal Curve (parameterized nonlinear manifold.

### 2.1 Linear PCA Manifolds

In PCA [18], the basis functions in a discrete Karhunen-Loève Transform (KLT) [22] are obtained by solving the algebraic eigenvalue problem $\mathbf{\Lambda} = \mathbf{\Phi}^T \mathbf{\Sigma} \mathbf{\Phi}$, where $\mathbf{\Sigma}$ is the covariance matrix of the data, $\mathbf{\Phi}$ is the eigenvector matrix of $\mathbf{\Sigma}$, and $\mathbf{\Lambda}$ is the corresponding diagonal matrix of eigenvalues. The unitary matrix $\mathbf{\Phi}$ defines a coordinate transform (rotation) which *decorrelates* the data and makes explicit the *invariant subspace* of the matrix "operator" $\Sigma$. Most commonly, PCA is a partial KLT which identifies the largest (or principal) eigenvalue eigenvectors for projecting the data: $\mathbf{y} = \mathbf{\Phi}_M^T \mathbf{x}$, where $\mathbf{\Phi}_\mathrm{M}$ is a submatrix of $\mathbf{\Phi}$ containing the principal eigenvectors (from here on, we will just use $\mathbf{\Phi}$ to denote $\mathbf{\Phi}_M$). PCA can be seen as a linear projection $\mathcal{R}^N \rightarrow \mathcal{R}^M$ onto the lower-dimensional subspace corresponding to the maximal eigenvalues. The main properties of the PCA transform are summarized by the following:

$$\mathbf{x} \approx \mathbf{\Phi} \mathbf{y},\ \mathbf{\Phi}^T \mathbf{\Phi} = \mathbf{I},\ E\{y_i y_j\}_{i \neq j} = 0 \tag{1}$$

corresponding to approximate reconstruction, orthonormality of the basis $\mathbf{\Phi}$ and decorrelated principal components, respectively. Fig. 1a illustrates the principal component vectors (columns of $\mathbf{\Phi}$) obtained with a toy data set corresponding to an essentially one-dimensional (nonlinear) manifold. Projection of the data points onto the first principal component would then correspond a 1D linear manifold representation (the second PC, shown as a smaller line segment in the figure, would be discarded in this low-dimensional example).

### 2.2 Linear ICA Manifolds

Independent Component Analysis (ICA) [20], [9] is similar to PCA except that the distribution of the components are designed to be sub/super Gaussian (usually by minimizing/maximizing fourth-order distribution cumulants such as kurtosis). Maximizing non-Gaussianity also promotes statistical *independence*, which is the desired goal. Like PCA, ICA is also a linear projection $\mathcal{R}^N \rightarrow \mathcal{R}^M$ but with different properties:

$$\mathbf{x} \approx \mathbf{A} \mathbf{y}, \mathbf{A}^T \mathbf{A} \neq \mathbf{I},\ P(\mathbf{y}) \approx \prod\ p(y_i) \tag{2}$$

corresponding to approximate reconstruction, *nonorthogonality* of the basis $\mathbf{A}$ and the near factorization of the joint distribution $P(\mathbf{y})$ into marginal distributions of the (non-Gaussian) ICs. An example of an ICA basis is shown in Fig. 1b, where we see two unordered nonorthogonal IC vectors one of which is roughly aligned with the first principal component vector in Fig. 1a—i.e., the direction of maximum variance. The actual non-Gaussianity and statistical independence achieved in this toy example are minimal at best.

### 2.3 Nonlinear Principal Manifolds

The defining property of nonlinear principal manifolds is that the *inverse image* of the manifold in the original space $\mathcal{R}^N$ is (typically) a nonlinear (curved) lower-dimensional surface that "passes through the middle of the data," while minimizing the sum total distance between the data points and their projections on that surface. Often referred to as *principal curves* [16], this formulation is essentially a nonlinear regression on the data. An example of a principal curve is shown in Fig. 1c.

One of the simplest methods for computing nonlinear principal manifolds is the nonlinear PCA (NLPCA) autoencoder multilayer neural network [21], [12] shown in Fig. 2. The so-called "bottleneck" layer forms a lower-dimensional manifold representation by means of a (weighted-sum-of-sigmoids) nonlinear *projection* function $f(\mathbf{x})$. The resulting principal components $\mathbf{y}$ have an inverse mapping with a similar nonlinear *reconstruction* function $g(\mathbf{y})$, which reproduces the input data as accurately as possible. The NLPCA computed by such a multilayer sigmoidal neural network is equivalent—with certain exceptions[2]—to a *principal surface* under the more general definition [15], [16]. To summarize, the main properties of NLPCA are:

$$\mathbf{y} = f(\mathbf{x}), \mathbf{x} \approx g(\mathbf{y}),\ P(\mathbf{y}) = ? \tag{3}$$

corresponding to nonlinear projection, approximate reconstruction and, typically, no prior knowledge regarding the joint distribution of the components, respectively (however, see Zemel and Hinton [45] for an example of devising suitable priors in such cases). The principal curve in Fig. 1c

2. The class of functions attainable by this neural network restricts the projection function $f()$ to be smooth and differentiable, hence suboptimal in some cases [24].

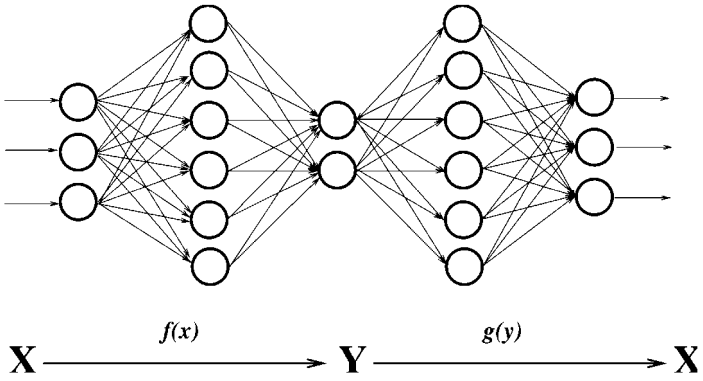



Fig. 2. An autoassociative ("bottleneck") neural network for computing principal manifolds $\mathbf{y} \in \mathcal{R}^M$ in the input space $\mathbf{x} \in \mathcal{R}^N$.

was generated with a 2-4-1-4-2 layer neural network of the type shown in Fig. 2. Note how the principal curve yields a compact and (relatively) accurate representation of the data.

### 2.4 Kernel PCA Manifolds

Recently, nonlinear principal component analysis has been revived with the "kernel eigenvalue" method of Schölkopf et al. [40]. The basic methodology of KPCA is to apply a nonlinear mapping to the input $\Psi(\mathbf{x}) : \mathcal{R}^N \rightarrow \mathcal{R}^L$ and then solve for a linear PCA in the resulting feature space $\mathcal{R}^L$, where $L$ is larger than $N$ and possibly infinite. Because of this increase in dimensionality, the mapping $\Psi(\mathbf{x})$ is made implicit (and economical) by the use of kernel functions satisfying Mercer's theorem [11]

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\Psi(\mathbf{x_i}) \cdot \Psi(\mathbf{x_j})), \tag{4}$$

where kernel evaluations $k(\mathbf{x_i}, \mathbf{x_j})$ in the input space correspond to dot-products in the higher dimensional feature space. Because computing covariance is based on dot-products, performing a PCA in the feature space can be formulated with kernels in the input space without the explicit (and possibly prohibitive) direct computation of $\Psi(\mathbf{x})$. Specifically, assuming that the projection of the data in feature space is zero-mean ("centered"), the covariance is given by

$$\Sigma_K = < \Psi(\mathbf{x_i}) \Psi(\mathbf{x_i})^T > \tag{5}$$

with the resulting eigenvector equation $\lambda \mathbf{V} = \Sigma_K \mathbf{V}$. Since the eigenvector solutions $\mathbf{V}$ must lie in the span of the training data $\Psi(\mathbf{x_i})$, it must be true that for each training point

$$\lambda(\Psi(\mathbf{x_i}) \cdot \mathbf{V}) = (\Psi(\mathbf{x_i}) \cdot \Sigma_K \mathbf{V}) \quad \text{for } i = 1, \ldots T \tag{6}$$

and that there must exist coefficients $\{w_i\}$ such that

$$\mathbf{V} = \sum_{i=1}^{T} w_i \Psi(\mathbf{x_i}). \tag{7}$$

Using the definition of $\Sigma_K$, substituting the above equation into (6), and defining the resulting $T$-by-$T$ matrix $\mathbf{K}$ by $\mathbf{K}_{ij} = (\Psi(\mathbf{x_i}) \cdot \Psi(\mathbf{x_j}))$, leads to the equivalent eigenvalue problem formulated in terms of kernels in the input space

$$T\lambda \mathbf{w} = \mathbf{K}\mathbf{w}, \tag{8}$$

where $\mathbf{w} = (w_1, \ldots, w_T)^T$ is the vector of expansion coefficients of a given eigenvector $\mathbf{V}$ as defined in (7). The kernel matrix $\mathbf{K}_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$ is then diagonalized with a standard PCA.[3] Orthonormality of the eigenvectors, $(\mathbf{V}^n \cdot \mathbf{V}^n) = 1$, leads to the equivalent normalization of their respective expansion coefficients, $\lambda_n(\mathbf{w}^n \cdot \mathbf{w}^n) = 1$.

Subsequently, the KPCA principal components of any input vector can be efficiently computed with simple kernel evaluations against the data set. The $n$th principal component $y_n$ of $\mathbf{x}$ is given by

$$y_n = (\mathbf{V}^n \cdot \Psi(\mathbf{x})) = \sum_{i=1}^{T} w_i^n k(\mathbf{x}, \mathbf{x_i}), \tag{9}$$

where $\mathbf{V}^n$ is the $n$th eigenvector of the feature space defined by $\Psi$. As with PCA, the eigenvectors $\mathbf{V}^n$ can be ranked by decreasing order of their eigenvalues $\lambda_n$ and an $M$-dimensional manifold projection of $\mathbf{x}$ is $\mathbf{y} = (y_1, \ldots, y_M)^T$, with individual components defined by (9).

A significant advantage of KPCA over neural network and principal curves is that KPCA does not require nonlinear optimization, is not subject to overfitting, and does not require prior knowledge of network architecture or number of dimensions. Furthermore, unlike traditional PCA, one can use more eigenvector projections than the input dimensionality of the data (since KPCA is based on the matrix $\mathbf{K}$, the number of eigenvectors or features available is $T$). On the other hand, the selection of the optimal kernel (and its associated parameters) remains an "engineering problem." Typical kernels include Gaussians $\exp(-\|\mathbf{x_i} - \mathbf{x_j}\|)^2/\sigma^2)$, polynomials $(\mathbf{x_i} \cdot \mathbf{x_j})^d$, and sigmoids $\tanh(a(\mathbf{x_i} \cdot \mathbf{x_j}) + b)$, all of which satisfy Mercer's theorem [11].

## 3 Probabilistic Subspaces

The input visual data (or equivalently its manifold representation) can form the basis for simple recognition strategies using Euclidean metrics or normalized correlation. For example, in its simplest form, the similarity measure $S(I_1, I_2)$ between two images $I_1$ and $I_2$ (or their manifold projections) can be set to be inversely proportional to the norm $||I_1 - I_2||$ which corresponds to a template-matching approach to recognition [6], [19]. Such a formulation suffers from a major drawback: it does not exploit knowledge of which types of variation are critical (as opposed to incidental)

3. However, computing $\Sigma_K$ in (5) requires "centering" the data by computing the mean of $\Psi(\mathbf{x_i})$. Since there is no explicit computation of $\Psi(\mathbf{x_i})$, the equivalent must be carried out when computing the kernel matrix $\mathbf{K}$. For details on "centering" $\mathbf{K}$, see [40].

in expressing similarity. However, one can formulate a *probabilistic* similarity measure which is based on the probability that the image intensity (or equivalently manifold vector) differences, denoted by $\Delta = I_1 - I_2$, are characteristic of typical variations in appearance of the *same* object. For example, for purposes of face recognition, one can define two classes of facial image variations: *intrapersonal* variations $\Omega_I$ (corresponding, for example, to different facial expressions of the *same* individual) and *extrapersonal* variations $\Omega_E$ (corresponding to variations between *different* individuals). The similarity measure $S(\Delta)$ can then be expressed in terms of the intrapersonal a posteriori probability given by Bayes rule:

$$S(\Delta) = P(\Omega_I|\Delta) = \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)}. \quad (10)$$

The likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$ can be estimated by traditional means (given enough data) or, alternatively, with subspace density estimation techniques [30], [26] when faced with very high-dimensional data or with data shortage (insufficient number of samples). Furthermore, the priors $P(\Omega)$ can be set to reflect specific operating conditions (e.g., number of test images versus the size of the database) or other sources of a priori knowledge regarding the two images being matched. Note that this particular Bayesian formulation casts the standard face recognition task (essentially an $\mathbf{m}$-ary classification problem for $\mathbf{m}$ individuals) into a *binary* pattern classification problem with $\Omega_I$ and $\Omega_E$. 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 $P(\Omega_I|\Delta) > P(\Omega_E|\Delta)$ or, equivalently, if $S(\Delta) > \frac{1}{2}$.

Alternatively, a simplified similarity measure based only on the $\Omega_I$ likelihood can be used. This *maximum-likelihood* (ML) similarity measure ignores extrapersonal variations altogether and is given by $S'(\Delta) = P(\Delta|\Omega_I)$. In our experience, the $\Omega_I$ density in (10) carries the greater weight in modeling the posterior similarity used for MAP recognition. The extrapersonal $\Omega_E$ density serves a secondary role and its accurate modeling is less critical. In the extreme, by dropping the $\Omega_E$ likelihood in favor of a ML similarity, we obtain $S'(\Delta)$, which typically suffers only a minor deficit (3-4 percent) in accuracy as compared to $S(\Delta)$ [27].

### 3.1 Subspace Density Estimation

To deal with the inevitable high-dimensionality of $\Delta$ (which is the same as that of the images), we use the efficient density estimation method proposed by Moghaddam and Pentland [29], [30] which divides the vector space $\mathcal{R}^N$ into two complementary subspaces as shown in Fig. 4 using an eigenspace decomposition. This method uses PCA to obtain a principal subspace $F$ whose principal components $\mathbf{y}$ can be used to form an optimal (minimal divergence) low-dimensional estimate of the complete likelihood using only the first $M$ principal components $\{y_1, y_2, y_3, \ldots, y_M\}$, where $M << N$.

As derived in [29], the complete likelihood estimate can be written as the product of two independent marginal Gaussian densities

$$\hat{P}(\Delta|\Omega) = \left[\frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{y_i^2}{\lambda_i}\right)}{(2\pi)^{M/2}\prod_{i=1}^{M}\lambda_i^{1/2}}\right] \cdot \left[\frac{\exp\left(-\frac{\epsilon^2(\Delta)}{2\rho}\right)}{(2\pi\rho)^{(N-M)/2}}\right] \quad (11)$$
$$= P_F(\Delta|\Omega)\hat{P}_{\bar{F}}(\Delta|\Omega;\rho),$$

where $P_F(\Delta|\Omega)$ is the true marginal density in $F$, $\hat{P}_{\bar{F}}(\Delta|\Omega;\rho)$ is the estimated marginal density in the orthogonal complement $\bar{F}$, $y_i$ are the principal components, and $\epsilon^2(\Delta)$ is the PCA residual (reconstruction error). The information-theoretic optimal value for the density parameter $\rho$ is derived by minimizing the Kullback-Leibler (KL) divergence and is found to be simply the average of the $\bar{F}$ eigenvalues

$$\rho = \frac{1}{N-M}\sum_{i=M+1}^{N}\lambda_i. \quad (12)$$

This derivation is a special case of a more recent and general factor analysis model, called Probabilistic PCA (PPCA), proposed by Tipping and Bishop [42]. In their formulation, the above expression for $\rho$ is the maximum-likelihood solution of a latent variable model as opposed to the minimal-divergence solution derived in [29]. For a more general expectation-maximization (EM) approach to factor analysis, the reader is referred to [39].

In actual practice, the majority of the $\bar{F}$ eigenvalues are unknown but can be estimated, for example, by fitting a nonlinear function to the available portion of the eigenvalue spectrum and estimating the average of the eigenvalues beyond the principal subspace. Fractal power law spectra of the form $f^{-n}$ are thought to be typical of "natural" phenomenon and are often a good fit to the decaying nature of the eigenspectrum—see Fig. 4b.

Referring back to (10), we see that this approach requires two projections of the difference vector $\Delta$, from which likelihoods can be estimated for the Bayesian similarity measure $S(\Delta)$. The projection steps are linear, while the posterior computation is nonlinear. Because of the double PCA projections required, this approach has been called a "dual eigenspace" technique [31], [26], [27] and is contrasted to standard PCA-based "eigenfaces" in Fig. 3. Note the projection of the difference vector $\Delta$ onto the "dual eigenfaces" ($\Omega_I$ and $\Omega_E$) for computation of the posterior in (10). In the following section, we will show that, in practice, each input vector $\mathbf{x}$ will have two (precomputed) linear PCA projections $\mathbf{y}_{\Phi_I}$ and $\mathbf{y}_{\Phi_E}$ and that the posterior similarity $S(\Delta)$ between any pair of vectors can be expressed in terms of a pair of difference norms between their corresponding dual projections.

### 3.2 Efficient Similarity Computation

Consider a feature space of $\Delta$ vectors, the differences between two images ($I_j$ and $I_k$). The two classes of interest in this space correspond to intrapersonal and extrapersonal variations and each is modeled as a high-dimensional Gaussian density as in (13). The densities are zero-mean since for each $\Delta = I_j - I_k$ there exists a $\Delta = I_k - I_j$.

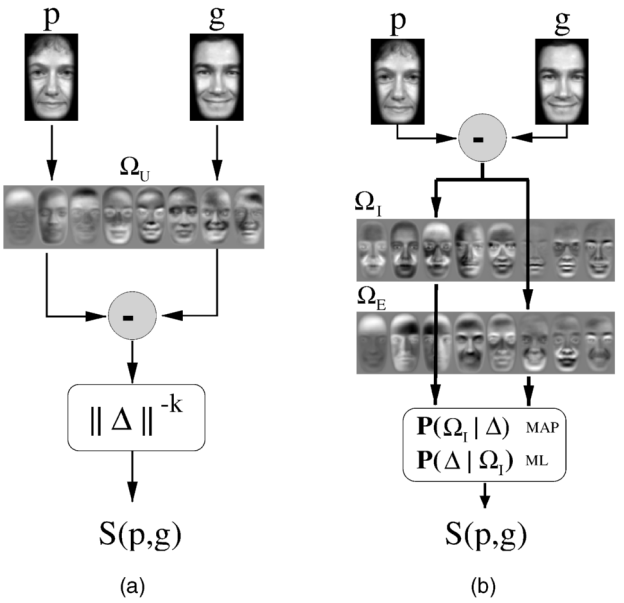



Fig. 3. Signal flow diagrams for computing similarity g between two images: (a) Eigenface similarity and (b) Probabilistic similarity. The difference image is projected through both sets of (intra/extra) eigenfaces in order to obtain the two likelihoods.

$$P(\Delta|\Omega_E) = \frac{e^{-\frac{1}{2}\Delta^T \Sigma_E^{-1} \Delta}}{(2\pi)^{D/2} |\Sigma_E|^{1/2}}$$
$$P(\Delta|\Omega_I) = \frac{e^{-\frac{1}{2}\Delta^T \Sigma_I^{-1} \Delta}}{(2\pi)^{D/2} |\Sigma_I|^{1/2}}. \quad (13)$$

By PCA, the Gaussians are known to only occupy a subspace of image space (face-space) and, thus, only the top few eigenvectors of the Gaussian densities are relevant for modeling. These densities are used to evaluate the similarity in (10). Computing the $S(\Delta)$ similarity involves first subtracting a candidate image $I_j$ from a database entry $I_k$. The resulting $\Delta$ image is then projected onto the eigenvectors of the extrapersonal Gaussian and also the eigenvectors of the intrapersonal Gaussian. The exponentials are computed, normalized, and then combined as in (10). This operation is iterated over all members of the database (many $I_k$ images) until the maximum score is found (i.e., the match). Thus, for large databases, such evaluations are expensive and must be simplified by offline transformations.

To compute the likelihoods $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$, we preprocess the $I_k$ images with whitening transformations. Each image is converted and stored as a set of two whitened subspace coefficients, $\mathbf{y}_{\Phi_I}$ for intrapersonal space and $\mathbf{y}_{\Phi_E}$ for extrapersonal space (see (14)). Here, $\Lambda$ and $V$ are matrices of the largest eigenvalues and eigenvectors of $\Sigma_E$ or $\Sigma_I$.

$$\mathbf{y}^j_{\Phi_I} = \Lambda_I^{-\frac{1}{2}} V_I I_j \qquad \mathbf{y}^j_{\Phi_E} = \Lambda_E^{-\frac{1}{2}} V_E I_j. \quad (14)$$

After this preprocessing, evaluating the Gaussians can be reduced to simple Euclidean distances as in (15). Denominators are of course precomputed. These likelihoods are evaluated and used to compute the MAP similarity $S(\Delta)$ in (10). Euclidean distances are computed between the $M_E$-dimensional $\mathbf{y}_{\Phi_I}$ vectors, as well as the $M_E$-dimensional $\mathbf{y}_{\Phi_E}$ vectors. Thus, roughly $2 \times (M_E + M_I)$ arithmetic operations are required for each similarity computation, avoiding repeated image differencing and projections.

$$P(\Delta|\Omega_I) = P(I_j - I_k|\Omega_I) = \frac{e^{-\|\mathbf{y}^j_{\Phi_I} - \mathbf{y}^k_{\Phi_I}\|^2/2}}{(2\pi)^{D/2} |\Sigma_I|^{1/2}}$$
$$P(\Delta|\Omega_E) = P(I_j - I_k|\Omega_E) = \frac{e^{-\|\mathbf{y}^j_{\Phi_E} - \mathbf{y}^k_{\Phi_E}\|^2/2}}{(2\pi)^{D/2} |\Sigma_E|^{1/2}}. \quad (15)$$

The ML similarity matching is even simpler since only the intrapersonal class is evaluated, leading to the following modified form for the similarity measure

$$S'(\Delta) = P(\Delta|\Omega_I) = \frac{e^{-\|\mathbf{y}^j_{\Phi_I} - \mathbf{y}^k_{\Phi_I}\|^2/2}}{(2\pi)^{D/2} |\Sigma_I|^{1/2}}. \quad (16)$$

## 4 Experiments

Our experimental data consisted of a training "gallery" of 706 individual FERET faces and 1,123 "probe" images containing one or more views of every person in the gallery. All these images were aligned and normalized, as described in [30]. The multiple probe images reflected different expressions, lighting, and with glasses on/off, etc. In this study, we decided to test the limits of the Bayesian matching algorithm

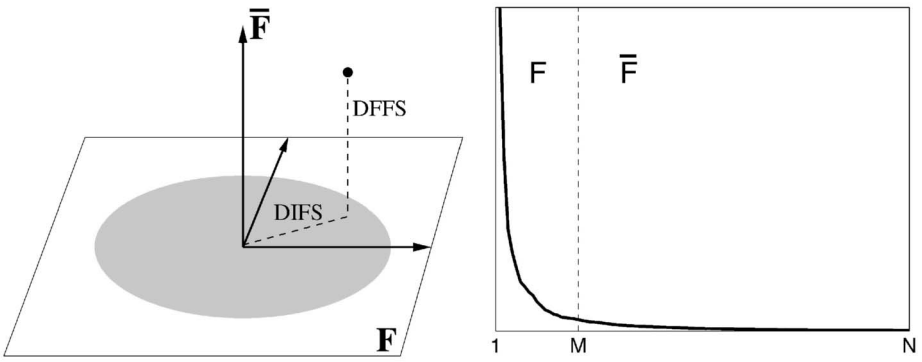
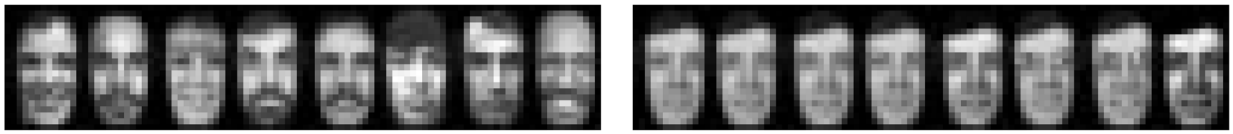
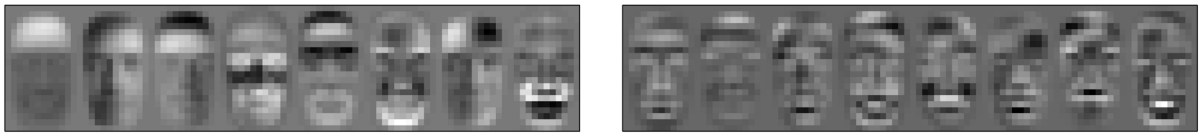



Fig. 4. (a) Decomposition of $\mathcal{R}^N$ into the principal subspace $F$ and its orthogonal complement $\bar{F}$ for a Gaussian density and (b) a typical eigenvalue spectrum and its division into the two orthogonal subspaces.



(a) (b)

Fig. 5. (a) Several faces from the gallery. (b) Multiple probes for one individual, with different facial expressions, eye-glasses, variable ambient lighting, and image contrast, etc.



(a) (b)

Fig. 6. (a) PCA faces and (b) ICA faces.

with respect to image resolution or, equivalently, the amount of visible facial detail. Since this algorithm was independently evaluated in DARPA's 1996 FERET face recognition competition [38] with medium resolution images (84-by-44 pixels)—achieving an accuracy of $\approx 95$ percent on $O(10^3)$ individuals—we decided to lower the resolution by a factor 16. Therefore, the aligned faces in our data set were downsampled to 21-by-12 pixels, yielding input vectors in a $\mathcal{R}^{N=252}$ space. Several examples are shown in Fig. 5.

### 4.1 Evaluation Methodology

In order to make maximum use of the available data for testing and to obtain confidence intervals on the recognition rates, we used a 5-fold Cross-Validation (CV) analysis. The total data set of 1,829 faces (706 unique individuals and their collective 1,123 probes) was randomly partitioned into five subsets with unique (nonoverlapping) individuals and their associated probes. Each subset contained both gallery and probe images of $\approx 140$ different (unique) individuals. For each of the five subsets, the recognition task was correctly matching the multiple probes to the $\approx 140$ gallery faces using the other four subsets as training data. Note that with $N = 252$ and using 80 percent of the entire data set for training, we had nearly three times as many training samples than the data dimensionality, thus parameter estimations (for PCA, ICA, KPCA, and Bayes) were properly overconstrained.

The resulting five experimental trials were pooled to compute the mean and standard deviation of the recognition rates for each method. The fact that the training and testing sets had no overlap in terms of individual identities leads to an evaluation of the algorithms' *generalization* performance—the ability to recognize new individuals which were not part of the manifold computation or density modeling with the training set.

For our baseline recognition experiments, we selected a default manifold dimensionality of $d = 20$. This choice of $d$ was made for two reasons: it led to a reasonable PCA reconstruction error of MSE = 0.0012 (or 0.12 percent per pixel with a normalized intensity range of [0,1]) and a baseline PCA recognition rate of $\approx 80$ percent (on a different 50/50 partition of the data set), thus leaving a sizeable margin for improvement. Note that since the recognition experiments were essentially a 140-way classification task, chance performance was approximately 0.7 percent.

#### 4.1.1 PCA-Based Recognition

The baseline algorithm for our face recognition experiments was standard PCA or "eigenface" matching. The first eight principal eigenvectors computed from a single partition are shown in Fig. 6a. Projection of the test set probes onto the 20-dimensional linear manifold (computed with PCA on the training set only) followed by nearest-neighbor matching to the $\approx 140$ gallery images using a Euclidean metric yielded a mean recognition rate of 77.31 percent with the highest rate achieved being 79.62 percent, as shown in Table 1. For calibration purposes, we also did full image-vector nearest-neighbor (template matching)—i.e., on $\mathbf{x} \in \mathcal{R}^{252}$—yielding a recognition rate of 86.46 percent (see dashed line in Fig. 7). Clearly, performance is degraded by the $252 \rightarrow 20$ dimensionality reduction, as expected.

#### 4.1.2 ICA-Based Recognition

For ICA-based recognition, we used two different algorithms based on fourth-order cumulants: the "JADE" algorithm of Cardoso [8] and the fixed-point algorithm of Hyvärinen and Oja [17]. In both algorithms, a PCA whitening step ("sphering") preceded the core ICA decomposition. The corresponding *nonorthogonal* JADE-derived ICA basis is shown in Fig. 6b. Similar basis faces were obtained with Hyvärinen's method. These basis faces are the columns of the matrix $\mathbf{A}$ in (2) and their linear combination (specified by the ICs) reconstructs the training data. The ICA manifold

TABLE 1
Recognition Accuracies (in Percent) with $d = 20$ Subspace Projections Using 5-Fold Crossvalidation

| Partition | PCA | ICA | KPCA | Bayes |
|---|---|---|---|---|
| 1 | 78.00 | 82.90 | 83.26 | 95.46 |
| 2 | 79.62 | 77.29 | 92.37 | 97.87 |
| 3 | 78.59 | 79.19 | 88.52 | 94.49 |
| 4 | 76.39 | 82.84 | 85.96 | 92.90 |
| 5 | 73.96 | 64.29 | 86.57 | 93.45 |
| Mean | 77.31 | 77.30 | 87.34 | 94.83 |
| Std. Dev. | 2.21 | 7.66 | 3.39 | 1.96 |

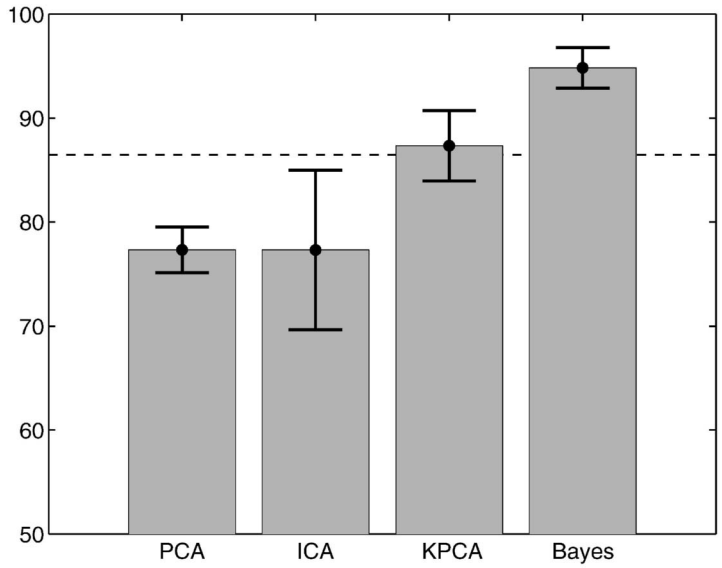



Fig. 7. Recognition performance of PCA, ICA, and KPCA manifolds versus Bayesian (MAP) similarity matching with a $d = 20$-dimensional subspace (dashed line indicates the performance of nearest-neighbor matching with the full-dimensional image vectors).

projection of the test set was obtained using $\mathbf{y} = \mathbf{A}^{-1}\mathbf{x}$. Nearest-neighbor matching with ICA resulted in a mean recognition rate of 77.30 percent with the highest rate being 82.90 percent, as shown in Table 1. We found little difference between the two ICA algorithms and noted that ICA resulted in the largest performance variation in the five trials (7.66 percent standard deviation). Based on the mean recognition rates, it seems doubtful that ICA provides a systematic advantage over PCA, indicating that "more non-Gaussian" and/or "more independent" components did not result in a better manifold for *recognition* purposes with this data set. Note that the experimental results of Bartlett et al. [2] with FERET faces did favor ICA over PCA, but mostly with more difficult time-separated images. Their ICA versus PCA performance margin at the $\approx 80$ percent recognition level was not as significant. Compared to Bartlett et al. [2] our faces were cropped much tighter leaving no information regarding hair and face shape and also were much lower in resolution; factors which when combined make the recognition task much harder.

#### 4.1.3 KPCA-Based Recognition

For KPCA, we first evaluated Gaussian, polynomial, and sigmoidal kernels and fine-tuned their parameters for best performance with a different 50/50 partition validation set—Gaussian kernels were found to be the best for this data set. For each trial, the kernel matrix was computed from the corresponding training data. Both the test set gallery and probes were projected onto the kernel eigenvector basis (using (9)) in order to obtain the nonlinear principal components which were then used in nearest-neighbor matching of test set probes against the test set gallery images. The mean recognition rate was found to be 87.34 percent with the highest rate being 92.37 percent, as shown in Table 1. The standard deviation of the KPCA trials was slightly higher (3.39) than that of PCA (2.21) but Fig. 7 indicates that KPCA does in fact do better than both PCA and ICA, hence justifying the use of nonlinear feature extraction.

#### 4.1.4 MAP-Based Recognition

For Bayesian similarity matching, appropriate training $\Delta$s for the two classes $\Omega_I$ (Fig. 5b) and $\Omega_E$ (Fig. 5a) were used for the dual PCA-based density estimates $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$, which were both modeled as single Gaussians with subspace dimensions of $M_I$ and $M_E$, respectively. The total subspace dimensionality $d$ was divided evenly between the two densities by setting $M_I = M_E = d/2$ for modeling.[4]

With $d = 20$, we used Gaussian subspace dimensions of $M_I = 10$ and $M_E = 10$ for $P(\Delta|\Omega_I)$ and $P(\Delta|\Omega_E)$, respectively. Note that $M_I + M_E = 20$, thus matching the total number of projections used with the three principal manifold techniques. Using the maximum a posteriori (MAP) similarity in (10), the Bayesian matching technique yielded a mean recognition rate of 94.83 percent with the highest rate achieved being 97.87 percent, as shown in Table 1. The standard deviation of the five partitions for this algorithm was also the lowest (1.96)—see Fig. 7).

### 4.2 Compactness of Manifolds

We also compared the performance of the different methods with different size manifolds by plotting their recognition rates $R(d)$ as a function of the first $d$ principal components. For the manifold matching techniques, this simply meant using a subspace dimension of $d$ (the first $d$ components of PCA/ICA/KPCA), whereas for the Bayesian matching technique this meant that the subspace Gaussian dimensions should satisfy $M_I + M_E = d$. Thus, all methods used the same number of subspace projections. This test was the premise for one of the key points investigated in this study: given the *same* number of subspace projections, which of these techniques is better at data modeling and subsequent recognition? The presumption being that the one achieving the highest recognition rate with the smallest dimension is preferred.

For this particular dimensionality test, the total data set of 1,829 images was partitioned (split) in half: a training set of

4. In practice, $M_I > M_E$ often works just as well. In fact, as $M_E \to 0$, one obtains a maximum-likelihood similarity $S = P(\Delta|\Omega_I)$ with $M_I = d$, which for this data set is only few percent less accurate than MAP [27].

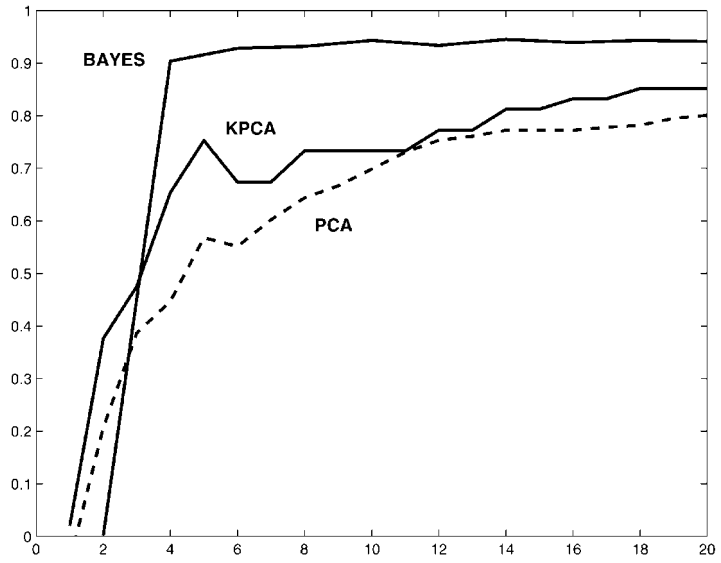



Fig. 8. Recognition accuracy $R(d)$ of PCA, KPCA, and Bayesian similarity with increasing dimensionality $d$ of the principal subspace (ICA results, not shown, are similar to PCA).

353 gallery images (randomly selected) along with their corresponding 594 probes and a testing set containing the remaining 353 gallery images and their corresponding 529 probes. The training and test sets had no overlap in terms of individuals' identities. As in the previous experiments, the test set probes were matched to the test set gallery images based on the projections (or densities) computed with the training set. The results of this experiment are shown in Fig. 8 which plots the recognition rates as a function of the dimensionality of the subspace $d$. This is a more revealing comparison of the relative performance of the different methods since *compactness* of the manifolds—defined by the lowest acceptable value of $d$—is an important consideration in regards to both generalization error (overfitting) and computational requirements.

## 5 Discussion

The relative performance of the principal manifold techniques versus Bayesian matching is summarized in Table 1 and Fig. 7. The advantage of probabilistic matching over metric matching on both linear and nonlinear manifolds is quite evident ($\approx 18$ percent increase over PCA and $\approx 8$ percent over KPCA). Note that the dimensionality test results in Fig. 8 indicate that KPCA out-performs PCA by a $\approx 10$ percent margin and, even more so, with only few principal components (a similar effect is reported by Schölkopf [40] were KPCA out-performs PCA in low-dimensional manifolds). However, Bayesian matching achieves $\approx 90$ percent with only four projections—two for each $P(\Delta|\Omega)$—and dominates both PCA and KPCA throughout the entire range of subspace dimensions in Fig. 8.

A comparison of the subspace techniques with respect to multiple criteria is shown in Table 2. Note that PCA, KPCA, and the dual subspace density estimation are uniquely defined for a given training set (making experimental comparisons repeatable), whereas ICA is not unique due to the variety of different techniques used to compute the basis and the iterative (stochastic) optimizations involved. Considering the relative computation (of training), KPCA required $\approx 7 \times 10^9$ floating-point operations compared to PCA's $\approx 2 \times 10^8$. On the average, ICA computation was one order of magnitude larger than PCA. Since the Bayesian similarity method's learning stage involves two separate PCAs, its computation is merely twice that of PCA (essentially the same order of magnitude).

Considering its significant performance advantage (at low-subspace dimensionality) and its relative simplicity, the dual-eigenface Bayesian matching method is a highly effective subspace modeling technique for face recognition. In independent FERET tests conducted by the US Army Laboratory [38], the Bayesian similarity technique outperformed PCA and other subspace techniques such as Fisher's Linear Discriminant (by a margin of at least 10 percent). The new experimental results in this paper show that a similar recognition accuracy can be achieved using mere "thumbnails" which are 16 times lower in resolution than the images used in the FERET test. These results demonstrate the Bayesian matching technique's robustness with respect to image resolution, thus revealing the surprisingly small amount of facial detail that is required for high-accuracy performance with this learning technique.

## Acknowledgments

The author would like to thank the reviewers for their careful and constructive criticism of the original manuscript. The revision and additional experiments have significantly improved the quality of the work presented. The author would also like to thank Tony Jebara for helpful discussions and collaboration in the derivations in Section 3.2.

TABLE 2
Comparison of the Subspace Techniques Across Multiple Attributes ($d = 20$)

| | PCA | ICA | KPCA | Bayes |
|---|---|---|---|---|
| Accuracy | 77% | 77% | 87% | 95% |
| Computation | $10^8$ | $10^9$ | $10^9$ | $10^8$ |
| Uniqueness | yes | no | yes | yes |
| Projections | linear | linear | nonlinear | linear |

## REFERENCES


[1] H. Attias, "Independent Factor Analysis," *Neural Computation,* vol. 11, no. 4, pp. 803-851, 1999.

[2] M.S. Bartlett, H.M. Lades, and T.J. Sejnowski, "Independent Component Representations for Face Recognition," *Proc. SPIE, Conf. Human Vision and Electronic Imaging III,* vol. 2399, pp. 528-539, 1998.

[3] V.I. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[4] C.M. Bishop, "Bayesian PCA," *Advances in Neural Information Processing Systems,* pp. 482-388, MIT Press, 1999.

[5] C. Bregler and S.M. Omohundro, "Surface Learning with Applications to Lip Reading," *Advances in Neural Information Processing Systems 6,* pp. 43-50, 1994.

[6] R. Brunelli and T. Poggio, "Face Recognition: Features vs. Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 10, Oct. 1993.

[7] M.C. Burl, U.M. Fayyad, P. Perona, P. Smyth, and M.P. Burl, "Automating the Hunt for Volcanos on Venus," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1994.

[8] J.-F. Cardoso, "High-Order Contrasts for Independent Component Analysis," *Neural Computation,* vol. 11, no. 1, pp. 157-192, 1999.

[9] P. Comon"Independent, Component Analysis—A New Concept?" *Signal Processing,* vol. 36, pp. 287-314, 1994.

[10] T.F. Cootes and C.J. Taylor, "Active Shape Models: Smart Snakes," *Proc. British Machine Vision Conf.,* pp. 9-18, 1992.

[11] R. Courant and D. Hilbert, *Methods of Mathematical Physiacs,* vol. 1. New York: Interscience, 1953.

[12] D. DeMers and G. Cottrell, "Nonlinear Dimensionality Reduction," *Advances in Neural Information Processing Systems 5,* 1993.

[13] K. Etemad and R. Chellappa, "Discriminant Analysis for Recognition of Human Faces," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 2148-2151, 1996.

[14] B.J. Frey and G.E. Hinton, "Variational Learning in Nonlinear Gaussian Belief Networks," *Neural Computation,* vol. 11, no. 1, pp. 193-213, 1999.

[15] T. Hastie, "Principal Curves and Surfaces," PhD thesis, Stanford Univ., 1984.

[16] T. Hastie and W. Stuetzle, "Principal curves," *J. Am. Statistical Assoc.,* vol. 84, no. 406, pp. 502-516, 1989.

[17] A. Hyvärinen and E. Oja, "A Family of Fixed-Point Algorithms for Independent Component Analysis," Technical Report A40, Helsinki Univ. of Technology, 1996.

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

[19] M.J. Jones and T. Poggio, "Model-Based Matching by Linear Combination of Prototypes," AI Memo No. 1583, Artificial Intelligence Laboratory, Massachusetts Inst. of Technology. Nov. 1996.

[20] C. Jutten and J. Herault, "Blind Separation of Sources," *Signal Processing,* vol. 24, pp. 1-10, 1991.

[21] M.A. Kramer, "Nonlinear Principal Components Analysis Using Autoassociative Neural Networks," *Am. Inst. Chemical Eng. J.,* vol. 32, no. 2, pp. 233-243, 1991.

[22] M.M. Loè, *Probability Theory.* Princeton: Van Nostrand, 1955.

[23] S. Makeig, A.J. Bell, T. Jung, and T.J. Sejnowski, "Independent Component Analysis of Electroencephalographic Data," *Advances in Neural Information Processing Systems 8,* pp. 145-151, 1996.

[24] E.C. Malthouse, "Some Theoretical Results on Nonlinear Principal Component Analysis," technical report, Northwestern Univ., 1998.

[25] M.J. McKeown, S. Makeig, T Jung, A.J. Bell, and T.J. Sejnowski, "Analysis of fMRI Data by Blind Separation into Spatial Independent Components," *Human Brain Mapping,* vol. 6, pp. 160-188, 1998.

[26] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian Modeling of Facial Similarity," *Advances in Neural Information Processing Systems 11,* pp. 910-916, 1998.

[27] B. Moghaddam, T. Jebara, and A. Pentland, "Efficient MAP/ML Similarity Matching for Face Recognition," *Proc. Int'l Conf. Pattern Recognition,* Aug. 1998.

[28] B. Moghaddam, C. Nastar, and A. Pentland, "Bayesian Face Recognition Using Deformable Intensity Differences," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR '96),* pp. 638-645, June 1996.

[29] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Detection," *Proc. Int'l Conf. Computer Vision (ICCV '95),* pp. 786-793, June 1995.

[30] B. Moghaddam and A. Pentland, "Probabilistic Visual Learning for Object Representation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 696-710, July 1997.

[31] B. Moghaddam, W. Wahid, and A. Pentland, "Beyond Eigenfaces: Probabilistic Matching for Face Recognition," *Proc. Int'l Conf. Automatic Face and Gesture Recognition (FG '98),* pp. 30-35, Apr. 1998.

[32] H. Murase and S.K. Nayar, "Learning and Recognition of 3D Objects from Appearance," *Proc. Qualitative Vision Workshop (CVRP '93),* June 1993.

[33] H. Murase and S. Nayar, "Visual Learning and Recognition of 3D Objects from Appearance," *Int'l J. Computer Vision,* vol. 14, no. 5, 1995.

[34] S.K. Nayar, S. Baker, and H. Murase, "Parametric Feature Detection," *Proc. EEE Conf. Computer Vision and Pattern Recognition,* pp. 471-477, June 1996.

[35] S.K. Nayar, H. Murase, and S.A. Nene, "General Learning Algorithm for Robot Vision," *Neural and Stochastic Methods in Image and Signal Processing,* vol. 2304, 1994.

[36] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* June 1994.

[37] A. Pentland and S. Sclaroff, "Closed-Form Solutions for Physically Based Shape Modeling and Recovery," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 13, no. 7, pp. 715-729, July 1991.

[38] P.J. Phillips, H. Moon, P. Rauss, and S. Rizvi, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Proc. Computer Vision and Pattern Recognition,* pp. 137-143, June 1997.

[39] D. Rubin and D. Thayer, "EM Algorithms for ML Factor Analysis," *Psychometrika,* vol. 47, no. 1, pp. 69-76, 1982.

[40] B. Schölkopf, A. Smola, and K.-R. Muller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. 10, no. 5, pp. 1299-1319, 1998.

[41] S. Sclaroff and A. Pentland, "Modal Matching for Correspondence and Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 6, pp. 545-561, June 1995.

[42] M. Tipping and C. Bishop, "Probabilistic Principal Component Analysis," Technical Report NCRG/97/010, Aston Univ., Sept. 1997.

[43] M. Tipping and C. Bishop, "Mixture of Probabilistic Principal Component Analyzers," *Neural Computation,* vol. 11, no. 2, pp. 443-482, 1999.

[44] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, 1991.

[45] R.S. Zemel and G.E. Hinton, "Developing Population Codes by Minimizing Description Length," *Advances in Neural Information Processing Systems,* J.D. Cowan, G. Tesauro, and J. Alspector, eds., vol. 6, Morgan Kaufmann, pp. 11-18, 1994.


**Baback Moghaddam** received the BS (magna cum laude) and MS (Honors) degrees in electrical and computer engineering from George Mason University in 1989 and 1992, respectively, and the PhD degree from the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT), in 1997. During his doctoral studies at MIT, he was a research assistant in the Vision and Modeling Group at the MIT Media Laboratory, where he developed an automatic face recognition system which was the top competitor in DARPA's "FERET" face recognition competition. Currently, he is a research scientist at Mitsubishi Electric Research Laboratory (MERL). Since joining MERL, Dr. Moghaddam has worked on visual demographic analysis for surveillance systems, Bayesian face recognition and fingerprint analysis for personal biometrics, user-centric image retrieval, and visualization of personal photo libraries and, most recently, on factorized density models of local structure and visual appearance for object representation. His general research interests include computer vision, image processing, computational learning theory, and statistical pattern recognition. Dr. Moghaddam is on the editorial board of the journal *Pattern Recognition*, has published more than 30 conference papers and 10 journal publications, and is a member of IEEE and ACM.