Appears in the *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **CVPR'98**, Santa Barbara, California, USA, June 23-25, 1998.

Probabilistic Reasoning Models for Face Recognition

Chengjun Liu and Harry Wechsler
Department of Computer Science, George Mason University,
4400 University Drive, Fairfax, VA 22030-4444, USA
{cliu, wechsler}@cs.gmu.edu

Abstract

We introduce in this paper two probabilistic reasoning models (PRM-1 and PRM-2) which combine the Principal Component Analysis (PCA) technique and the Bayes classifier and show their feasibility on the face recognition problem. The conditional probability density function for each class is modeled using the within class scatter and the Maximum A Posteriori (MAP) classification rule is implemented in the reduced PCA subspace. Experiments carried out using 1107 facial images corresponding to 369 subjects (with 169 subjects having duplicate images) from the FERET database show that the PRM approach compares favorably against the two well-known methods for face recognition — the Eigenfaces and Fisherfaces.

1. Introduction

A successful face recognition methodology depends heavily on the particular choice of the features used by the (pattern) classifier [3], [14], [2]. Feature selection in pattern recognition involves the derivation of salient features from the raw input data in order to reduce the amount of data used for classification and simultaneously provide enhanced discriminatory power. The selection of an appropriate set of features often exploits the design criteria such as (A) redundancy minimization and decorrelation, (B) minimization of the reconstruction error, (C) maximization of information transmission (infomax) [9], and (D) sparseness of the neural code [11]. The actual performance of a pattern recognition system depends, however, not only on the features to be chosen, but also on the classifier being used. The Bayes classifier yields the minimum error when the underlying probability density functions (pdf) are known. This error, called the Bayes error, is the optimal measure for feature effectiveness when classification is of concern, since it is a measure of class separability. The mean square error is an appropriate measure only when signal reconstruction rather than signal classification is of concern [5].

In this paper we describe two hybrid **P**robabilistic **R**easoning **M**odels (**PRM**) which combine the Principal Component Analysis (PCA) technique and the Bayes classifier, and show their feasibility on the face recognition problem. PCA is applied first as a dimensionality reduction technique with the goal of signal approximation. The PRM use then the within class scatter to estimate the covariance matrix for each class as required by the Bayes classifier to approximate the conditional pdf, and then apply the Maximum *A Posteriori* (MAP) as the classification criterion. The within class scatter and MAP decision rule optimize the class separability in the sense of Bayes error and should improve on PCA and Fisher Linear Discriminant (FLD) based methods, which utilize criteria not related to the Bayes error (Fukunaga [5]). Experiments carried out using 1107 facial images corresponding to 369 subjects (with 169 subjects having duplicate images) from the FERET database show that the PRM approach compares favorably against the two well-known face recognition methods — the Eigenfaces and Fisherfaces .

2. Background

Principal Component Analysis (PCA), also known as the Karhunen-Loeve expansion, is a classical technique for signal representation [7]. Sirovich and Kirby [8] applied PCA for representing face images. They showed that any particular face can be economically represented along the eigenpictures coordinate space, and that any face can be approximately reconstructed by using just a small collection of eigenpictures and the corresponding projections ('coefficients') along each eigenpicture.

Since eigenpictures are fairly good at representing face images, one could consider using the projections along them as classification features to recognize human faces. As a result, Turk and Pentland developed a well known face recognition method, known as **eigenfaces**, where the eigenfaces correspond to the eigenvectors associated with the dominant eigenvalues of the face covariance matrix. The eigenfaces define a feature space, or "face space", which drastically reduces the dimensionality of the original space, and face detection and identification are carried out in the reduced space [16].

The advantage of applying PCA directly for face recognition (eigenfaces) comes from its generalization ability [12]. PCA yields projection axes based on the variations from all the training samples, hence these axes are fairly robust for representing both training and testing images (not seen during training). The disadvantage of the direct PCA approach is that it does not distinguish the different roles of the variations, and treats all the variations equally. This will lead to poor performance when the distributions of the face classes are not separated by the mean-difference but separated by the covariance-difference [5].

While PCA is a classical technique for signal representation, Fisher's Linear Discriminant (FLD) is a classical technique for pattern recognition [4]. Several authors have applied this technique for face recognition, gesture recognition, and pattern rejection. As an example, Belhumire, Hespanha, and Kriegman [1] developed an approach called **Fisherfaces** by applying first PCA for dimensionality reduction and then FLD for discriminant analysis. Using a similar approach, Swets and Weng [15] have pointed out that the eigenfaces derived using PCA are only the most expressive features (MEF). The MEF are unrelated to actual face recognition, and in order to derive the most discriminating features (**MDF**), one needs a subsequent FLD projection. One can show that the MDF space is, however, superior to the MEF space for face recognition, only when the training images are representative of the range of face (class) variations; otherwise, the performance difference between the MEF and MDF is not significant.

The advantage of these indirect methods (combining PCA and FLD) is that they distinguish the different roles of within and between class scatter by applying discriminant analysis, e.g. FLD, and they usually produce non-orthogonal projection axes. But the indirect methods have their disadvantage too, namely poor generalization to new data, because those methods overfit to the training data. As the FLD procedure involves the simultaneous diagonalization of the two within and between class scatter matrices, it is equivalent to two-step operations: first 'whitening' the within class scatter matrix — applying an appropriate transformation that will make the within class scatter matrix equal to unity, and second applying PCA on the new between class scatter matrix [5]. The purpose of the 'whitening' step here is to normalize the within class scatter to unity, while the second step would then maximize the between class scatter. The robustness of the FLD procedure thus depends on whether or not the within class scatter can capture enough variations for a specific class. When the training samples do not include most of the variations due to lighting, facial expression, pose, and/or duplicate images as those encountered during testing, the 'whitening' step is likely to fit misleading variations, i.e. the normalized within class scatter would best fit the training samples but it would generalize poorly when exposed to new data. As a consequence the performances during testing for such an indirect method will deteriorate. In addition, when the training sample size for each class is small, the within class scatter would usually not capture enough variations. The FLD procedure thus leads to overfitting.

The **P**robabilistic **R**easoning **M**odels (**PRM**) detailed in the following section are based on Bayes classifier and the Maximum *A Posteriori* (MAP) criterion and they generalize many of the popular statistical face recognition methods: eigenfaces [16], PVL method [10], and Fisherfaces [1]. The PRM approach thus not only improves the recognition performance as compared to direct PCA approach (eigenfaces) and enhances the generalization capability of the indirect methods (Fisherfaces/MDF), but it also provides a unified statistical framework which allows formulation and comparison among these methods.

3. Probabilistic Reasoning Models (PRM)

The PRM work as follows. First PCA reduces the dimensionality of the original object space. The advantage of applying PCA comes from the merit of PCA as an optimal technique for signal representation through dimensionality reduction. PCA yields projection axes based on the variations from all the training samples, hence those axes are fairly robust for representing both training and testing images (not seen during training). The second step is to estimate the within class density and under the normal probability distribution assumption this step is equivalent to estimate the within class covariance matrices (within class scatter). Two PRM models are presented to estimate those matrices. One model (PRM-1) computes the within class scatter with respect to each principal component, under the assumption that all the within class covariance matrices are identical, diagonal. The other model (PRM-2) derives the within class scatter by computing the averaged within class covariance matrix based on all the within class scatters in the reduced PCA subspace, diagonalizing it, and utilizing the ordered diagonal elements as the estimation.

3.1. Dimensionality Reduction

The rationale behind applying PCA first for dimensionality reduction instead of exploiting the Bayes classifier and the MAP rule directly on the original data is two-fold; on the one hand, the high-dimensionality of the original face space makes the parameter estimation very difficult, if not impossible, due to the fact that high-dimensional space is mostly empty. This problem of sparsity limits the success of direct Bayesian analysis in the original space, since the amount of training data needed to get reasonably low variance estimators becomes ridiculously high [6]. On the other hand, it has been confirmed by many researchers that the PCA representation has the feature of object constancy in the sense that it suppresses input noise [8], [12].

PCA generates a set of orthonormal basis vectors, known as principal components, that maximize the scatter of all the projected samples. Let $X = [X_1, X_2, \ldots, X_n]$ be the sample set of the original images. After normalizing the images to unity norm and subtracting the grand mean a new image set $Y = [Y_1, Y_2, \ldots, Y_n]$ is derived. Each Y_i represents a normalized image with dimensionality N, $Y_i = (y_{i_1}, y_{i_2}, \ldots, y_{i_N})^t$, $(i = 1, 2, \ldots, n)$. The covariance matrix of the normalized image set is defined as

$$\Sigma_Y = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^t = \frac{1}{n} Y Y^t \tag{1}$$

and the eigenvector and eigenvalue matrices Φ, Λ are computed as

$$\Sigma_Y \Phi = \Phi \Lambda \tag{2}$$

Note that YY^t is an $N \times N$ matrix while Y^tY is an $n \times n$ matrix. If the sample size n is much smaller than the dimensionality N, then the following method saves some computation [16]

$$(Y^t Y)\Psi = \Psi \Lambda_1 \tag{3}$$

$$\Im = Y\Psi \tag{4}$$

where $\Lambda_1 = diag\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, and $\Im = [\Phi_1, \Phi_2, \ldots, \Phi_n]$. If one assumes that the eigenvalues are sorted in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, then the first m leading eigenvectors define matrix P

$$P = [\Phi_1, \Phi_2, \ldots, \Phi_m] \tag{5}$$

The new feature set Z with lower dimensionality m $(m \ll N)$ is then computed as

$$Z = P^t Y \tag{6}$$

3.2. The Bayes Classifier and MAP Classification Rule

For pattern recognition, the Bayes classifier is the best classifier, the Bayes error the best criterion to evaluate feature sets, and *a posteriori* probability functions are thus optimal features [5]. Let $\omega_1, \omega_2, \ldots, \omega_L$ denote the object classes, and Z an image in the reduced PCA subspace. The *a posteriori* probability function of ω_i given Z is defined as

$$P(\omega_i|Z) = \frac{p(Z|\omega_i)P(\omega_i)}{p(Z)} \tag{7}$$

where $P(\omega_i)$ is *a priori* probability, $p(Z|\omega_i)$ the conditional probability density function of ω_i, and $p(Z)$ is the mixture density. The Maximum *A Posteriori* (MAP) decision rule for the Bayes classifier is

$$p(Z|\omega_i)P(\omega_i) = \max_j \{p(Z|\omega_j)P(\omega_j)\}, \quad Z \in \omega_i \tag{8}$$

The face image Z is classified to ω_i of whom the *A Posteriori* probability given Z is the largest among all the classes.

Usually there are not enough samples to estimate the conditional probability density function for each class (within class density). A compromise, therefore, is to make an assumption of a particular density form, and convert the general density estimation question into a parametric one. The within class densities are usually modeled as normal distributions

$$\begin{aligned} p(Z|\omega_i) &= \frac{1}{(2\pi)^{m/2}|\Sigma_i|^{1/2}} \\ &\times \; exp\left\{-\frac{1}{2}(Z - M_i)^t \Sigma_i^{-1}(Z - M_i)\right\} \end{aligned} \tag{9}$$

where M_i (see Eq. 10) and Σ_i are the mean and covariance matrix of class ω_i, respectively.

3.3. A Unified Statistical Framework

Estimating the covariance matrix Σ_i in Eq. 9 with respect to each class is still difficult due to the limited number of samples for each class. Note that while the mixture covariance matrix is diagonal following PCA, the within class covariance matrices are not necessarily diagonal. Further assumptions lead to different face recognition methodologies:

<u>Case 1:</u> Assume the within class covariance matrices to be unit matrices: $\Sigma_I = \Sigma_i = I_m$, and under this assumption the conditional pdf (Eq. 9) relaxes to $p(Z|\omega_i) = \frac{1}{(2\pi)^{m/2}} exp\{-\frac{1}{2}(Z - M_i)^t(Z - M_i)\}$. As a result, the MAP rule (Eq. 8) leads to a distance classifier which corresponds to the Eigenfaces method by Turk and Pentland [16].

<u>Case 2:</u> By Modeling the conditional pdf of each class using the covariance matrix of all the samples corresponding to all classes and assuming the within class covariance matrices are identical, diagonal, and the same as the mixture covariance matrix, the MAP rule (Eq. 8) then specifies the Probabilistic Visual Learning (PVL) method by Moghaddam and Pentland [10]. In contrast to applying PCA primarily for dimensionality reduction, PVL method uses the eigenspace decomposition as an integral part of estimating complete density functions in high-dimensional image

spaces [10]. The first M eigenvalues are derived directly by PCA, and the remainder of the eigenvalue spectrum is estimated by curve fitting.

Case 3: After PCA (Eq. 6), apply FLD analysis to derive a new feature set $W = P_{FLD}^t Z$. In this FLD subspace, assume all the within class covariance matrices to be unit matrices: $\Sigma_I = \Sigma_i = I_m$, and under this assumption the conditional pdf (Eq. 9) reduces to $p(W|\omega_i) = \frac{1}{(2\pi)^{m/2}} exp\{-\frac{1}{2}(W - M_i')^t(W - M_i')\}$. Again the MAP rule (Eq. 8) leads to a distance classifier which corresponds to Fisherfaces method by Belhumeur, Hespanha, and Kriegman [1].

Case 4: In the PCA subspace (Eq. 6), assume all the within class covariance matrices are identical, diagonal: $\Sigma_I = \Sigma_i = diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\}$. Each diagonal element is estimated by sample variance in the one dimensional PCA subspace. As a result, the conditional pdf (Eq. 9) becomes $p(Z|\omega_i) = \frac{1}{(2\pi)^{m/2}\prod_{j=1}^m \sigma_j} exp\left\{-\frac{1}{2}\sum_{j=1}^m \frac{(z_j - m_{ij})^2}{\sigma_j^2}\right\}$ and it specifies the MAP rule (Eq. 8) as a quadratic classifier characterized by Mahalanobis distance. This case corresponds to our PRM-1 model.

Case 5: If we are reluctant to assume all the within class covariance matrices are diagonal and each diagonal element can be estimated by sample variance in the one dimensional PCA subspace, then we still can compute the averaged within class covariance matrix based on all the within class scatters in the reduced PCA subspace: $\Sigma_w = \frac{1}{L}\sum_{k=1}^L \left\{\frac{1}{N_k}\sum_{j=1}^{N_k}\left(Z_j^{(k)} - M_k\right)\left(Z_j^{(k)} - M_k\right)^t\right\}$. Diagonalize Σ_w, and use the ordered diagonal elements as estimations of the within class covariance matrices corresponding to those derived in case 4, and again the MAP rule (Eq. 8) specifies a quadratic classifier. This case corresponds to our PRM-2 model.

3.4. PRM-1 and PRM-2 Models

The two probabilistic reasoning models, PRM-1 and PRM-2, utilize the within class scatters to derive averaged estimations of within class covariance matrices. In particular, let $\omega_1, \omega_2, \ldots, \omega_L$ and $N_1, N_2, \ldots, N_L$ denote the classes and number of images within each class, respectively. Let $M_1, M_2, \ldots, M_L$ be the means of the classes in the reduced PCA subspace $span[\Phi_1, \Phi_2, \ldots, \Phi_m]$. We then have

$$M_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_j^{(i)}, \qquad i = 1, 2, \ldots, L \tag{10}$$

where $Z_j^{(i)}, j = 1, 2, \ldots, N_i$, represents the sample images from class ω_i.

The PRM-1 model assumes the within class covariance matrices are identical and diagonal (see Case 4)

$$\Sigma_I = \Sigma_i = diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\} \tag{11}$$

Each component σ_i^2 can be estimated by sample variance in the one dimensional PCA subspace

$$\sigma_i^2 = \frac{1}{L}\sum_{k=1}^L \left\{\frac{1}{N_k - 1}\sum_{j=1}^{N_k}\left(z_{ji}^{(k)} - m_{ki}\right)^2\right\} \tag{12}$$

where $z_{ji}^{(k)}$ is the i-th element of the sample $Z_j^{(k)}$, m_{ki} the i-th element of M_k, and L the number of classes.

From Eqs. 11 and 9, it follows

$$p(Z|\omega_i) = \frac{1}{(2\pi)^{m/2}\prod_{j=1}^m \sigma_j} exp\left\{-\frac{1}{2}\sum_{j=1}^m \frac{(z_j - m_{ij})^2}{\sigma_j^2}\right\} \tag{13}$$

Thus the MAP rule (Eq. 8) specifies a quadratic classifier characterized by Mahalanobis distance.

The PRM-2 model estimates the within class scatter matrix Σ_w in the reduced PCA subspace as

$$\Sigma_w = \frac{1}{L}\sum_{k=1}^L \left\{\frac{1}{N_k}\sum_{j=1}^{N_k}\left(Z_j^{(k)} - M_k\right)\left(Z_j^{(k)} - M_k\right)^t\right\} \tag{14}$$

To avoid the explicit calculation of Σ_w and to improve numerical accuracy, we calculate the singular value decomposition (SVD) of matrix Z, where $\Sigma_w = ZZ^t$.

$$Z = USV^t \tag{15}$$

where U and V are unitary matrices, S is a diagonal one

$$S = diag\{s_1, s_2, \ldots, s_m\} \tag{16}$$

with non-negative singular values as diagonal elements. Order the squared diagonal elements as

$$\left(s_{(1)}^2, s_{(2)}^2, \ldots, s_{(m)}^2\right) = order\{s_1^2, s_2^2, \ldots, s_m^2\} \tag{17}$$

Finally, the within class covariance matrix is derived as

$$\Sigma_I = diag\left\{s_{(1)}^2, s_{(2)}^2, \ldots, s_{(m)}^2\right\} \tag{18}$$

Under the assumption of Case 5, together with Eq. 9 and 8 the MAP rule specifies another quadratic classifier.

4. Experimental Results

After the estimations of the within class covariance matrices using Eqs. 11, 12 (for PRM-1) and Eq. 18 (for PRM-2), face recognition can be carried out using the Bayes decision rule (Eq. 8). The *a prior* probabilities are assigned values according to *a prior* knowledge, and are set to be equal in our experiments.

The experimental data consists of 1107 facial images corresponding to 369 subjects and comes from the US Army FERET database [13]. 600 out of the 1107 images correspond to 200 subjects with each subject having three images — two of them are the first and the second shot, and the third shot is taken under low illumination (see Fig. 1). For the remaining 169 subjects there are also three images for each subject, but two out of the three images are duplicates taken at a different time (see Fig. 1). Two images of each subject are used for training with the remaining image for testing. The images are cropped to the size of 64 x 96, and the eye coordinates are manually detected.

Figure 1. Examples of Face Images from FERET Database

Fig. 2 shows the comparative performances of Eigenfaces, Fisherfaces, PRM-1 and PRM-2 methods for the top 1 recognition rates. Top 1 recognition rate means the accuracy rate for the top response being correct, while top 3 recognition rate represents the accuracy rate for the correct response being included among the first three ranked choices. The top 3 recognition rates for the testing performance of the methods are plotted in Fig. 3. For the top 1 recognition rates, the PRM models (PRM-1 and PRM-2) increase the recognition rate by 5% when compared to Eigenfaces and Fisherfaces methods; the peak recognition rate for both PRMs is about 96% using 44 principal components. PRM-1 and PRM-2 models also increase the top 3 recognition rate by 3% when compared to Eigenfaces and Fisherfaces methods with a peak recognition rate of about 99%.

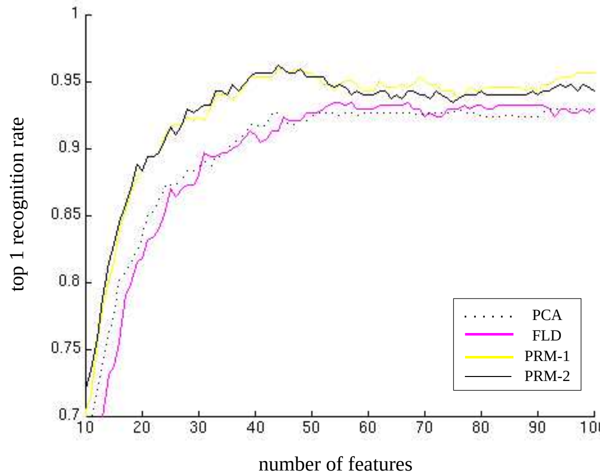

Figure 2. Comparative Testing Performances (top 1 recognition rate) for the PCA (Eigenfaces), FLD (Fisherfaces method), and PRM Approach (PRM-1, PRM-2).

5. Conclusions

We have introduced in this paper two probabilistic reasoning models (PRM-1 and PRM-2), which optimize the class separability in the sense of the Bayes error. Most related to our PRM method is the Probabilistic Visual Learning (**PVL**) method (Moghaddam and Pentland, [10]). Both PRM and PVL use a Bayesian probabilistic framework, but while PVL estimates the conditional densities of each class using the mixture class scatter PRM use the within class scatter which should enhance the discrimination ability of the classifier. Both methods apply PCA but with different purposes. PVL uses the eigenspace decomposition as an integral part of estimating complete density functions in high-dimensional image spaces [10]. The first M eigenvalues are derived directly by PCA, and the remainder of the eigenvalue spectrum is estimated by curve fitting. The PRM method, on the other hand, first applies PCA for dimensionality reduction and then exploits the Bayes classifier and the MAP rule for classification. The rationale behind applying PCA for dimensionality reduction is two-fold. First, the high-dimensionality of the original face space makes the parameter estimation very difficult, if not impossible, due to the fact that high-dimensional space is mostly empty [6]. Second, one should also point out that just using more PCs (principal components) does not necessarily lead to better performance, since the MAP rule (Eq. 8) involves a whitening factor (Eq. 9) which has a variance term in the denominator — the multiplication of this whitening factor ampli-

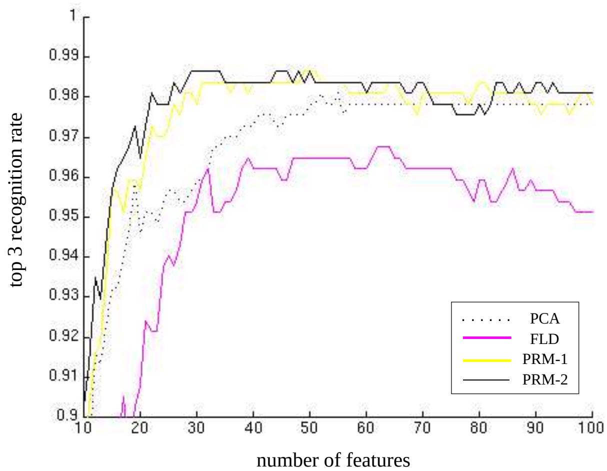

Figure 3. Comparative Testing Performances (top 3 recognition rate) for the PCA (Eigenfaces), FLD (Fisherfaces method), and PRM Approach (PRM-1, PRM-2).

fies both signal and noise, and while the power of the signal decreases with the increase of the index of PCs, the power of the noise remains constant. In order to fight noise amplification, the PRM method applies PCA for dimensionality reduction first, since it has been confirmed by many researchers that the PCA representation has the feature of object constancy in the sense that it suppresses input noise [8], [12]. Experimental results show that the PRM achieve better performance on face recognition when compared against both PCA and Fisherfaces methods. The PVL method is not experimented with because its performance depends on different functions used for curve fitting to estimate the remaining eigenvalues.

The next task for PRM is going to be choosing the best subset of principal components rather than the first ones corresponding to the leading eigenvalues. As an example, one can show that using more principal components (or eigenvalues) leads to decreased performance. The explanation for this behavior is that the ('high-frequency') eigenvalues in the later part of the whole spectrum usually capture mostly noise and their value is much smaller when compared against the leading eigenvalues. As a result, the Mahalanobis distance will artificially increase as the trailing eigenvalues appear in the denominator. There is also reason to believe that one could possibly delete the very first components as well in order to improve on performance. As PCA encode 2nd order statistics and are similar to the power spectrum one possible choice would be to weight the components with a filter approaching the known characteristics of the Contrast Sensitivity Function (CSF). Another possibility is to actually find the best subset of components using a greedy search type of algorithm.

Acknowledgments: This work was partially supported by the DoD Counterdrug Technology Development Program, with the U.S. Army Research Laboratory as Technical Agent, under contract DAAL01-97-K-0118.

References

[1] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[2] R. Brunelli and T. Poggio. Face recognition: Features vs. templates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(10):1042–1053, 1993.

[3] R. Chellappa, C. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proc. IEEE*, 83(5):705–740, 1995.

[4] R. Fisher. The use of multiple measures in taxonomic problems. *Ann. Eugenics*, 7:179–188, 1936.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1991.

[6] N. Intrator and L. Cooper. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17, 1992.

[7] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, 1986.

[8] 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):103–108, 1990.

[9] R. Linsker. Self-organization in a perceptual network. *Computer*, 21:105–117, 1988.

[10] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[11] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[12] P. Penev and J. Atick. Local feature analysis: A general statistical theory for object representation. *Network: Computation in Neural Systems*, 7:477–500, 1996.

[13] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET september 1996 database and evaluation procedure. In *Proc. First Int'l Conf. on Audio and Video-based Biometric Person Authentication*, pages 12–14, Switzerland, 1997.

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

[15] D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on PAMI*, 18(8):831–836, 1996.

[16] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 13(1):71–86, 1991.