Appears in the *Second International Conference on Audio- and Video-based Biometric Person Authentication*, **AVBPA'99**, Washington D. C. USA, March 22-24, 1999.

Comparative Assessment of Independent Component Analysis (ICA) 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

This paper addresses the relative usefulness of Independent Component Analysis (ICA) for Face Recognition. Comparative assessments are made regarding (i) ICA sensitivity to the dimension of the space where it is carried out, and (ii) ICA discriminant performance alone or when combined with other discriminant criteria such as Bayesian framework or Fisher's Linear Discriminant (FLD). Sensitivity analysis suggests that for enhanced performance ICA should be carried out in a compressed and whitened Principal Component Analysis (PCA) space where the small trailing eigenvalues are discarded. The reason for this finding is that during whitening the eigenvalues of the covariance matrix appear in the denominator and that the small trailing eigenvalues mostly encode noise. As a consequence the whitening component, if used in an uncompressed image space, would fit for misleading variations and thus generalize poorly to new data. Discriminant analysis shows that the ICA criterion, when carried out in the properly compressed and whitened space, performs better than the eigenfaces and Fisherfaces methods for face recognition, but its performance deteriorates when augmented by additional criteria such as the Maximum A Posteriori (MAP) rule of the Bayes classifier or the FLD. The reason for the last finding is that the Mahalanobis distance embedded in the MAP classifier duplicates to some extent the whitening component, while using FLD is counter to the independence criterion intrinsic to ICA.

1. Introduction

Face recognition is important not only because it has a lot of potential applications in research fields such as Human Computer Interaction (HCI), biometrics and security, but also because it is a typical Pattern Recognition (PR) problem whose solution would help solving other classification problems. A successful face recognition methodology depends heavily on the particular choice of the features used by the (pattern) classifier [4], [17], [3]. 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.

One popular technique for feature selection and dimensionality reduction is Principal Component Analysis (PCA) [8], [6]. PCA is a standard decorrelation technique and following its application one derives an orthogonal projection basis which directly leads to dimensionality reduction, and possibly to feature selection. PCA was first applied to reconstruct human faces by Kirby and Sirovich [10] and to recognize faces by Turk and Pentland [19]. The recognition method, known as eigenfaces, defines a feature space, or "face space", which drastically reduces the dimensionality of the original space, and face detection and identification are then carried out in the reduced space.

Independent Component Analysis (ICA) has emerged recently as one powerful solution to the problem of blind source separation [5], [9], [7] while its possible use for face recognition has been shown by Bartlett and Sejnowski [1]. ICA searches for a linear transformation to express a set of random variables as linear combinations of statistically independent source variables [5]. The search criterion involves the minimization of the mutual information expressed as a function of high order cumulants. Basically PCA considers the 2nd order moments only and it uncorrelates data, while ICA accounts for higher order statistics and it identifies the independent source components from their linear mixtures (the observables). ICA thus provides a more powerful data representation than PCA [9]. As PCA derives only the most expressive features for face reconstruction rather than face classification, one would usually use some subsequent discriminant analysis to enhance PCA performance [18].

This paper makes a comparative assessment on the use

of ICA as a discriminant analysis criterion whose goal is to enhance PCA stand alone performance. Experiments in support of our comparative assessment of ICA for face recognition are carried out using a large data set consisting of 1,107 images and drawn from the FERET database [16]. The comparative assessment suggests that for enhanced face recognition performance ICA should be carried out in a compressed and whitened space, and that ICA performance deteriorate when it is augmented by additional decision rules such as the Bayes classifier or the Fisher's linear discriminant analysis.

2. Background

PCA provides an optimal signal representation technique in the mean square error sense. The motivation behind using PCA for human face representation and recognition comes from its optimal and robust image compression and reconstruction capability [10] [15]. 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). PCA does not distinguish, however, the different roles of the variations (within- and between-class variations) and it treats them equally. This leads to poor performance when the distributions of the face classes are not separated by the mean-difference but separated by the covariance-difference [6].

Swets and Weng [18] point out that PCA derives only the most expressive features which are unrelated to actual face recognition, and in order to improve performance one needs additional discriminant analysis. One such discriminant criterion, the Bayes classifier, yields the minimum classification error when the underlying probability density functions (pdf) are known. The use of the Bayes classifier is conditioned on the availability of an adequate amount of representative training data in order to estimate the pdf. As an example, Moghaddam and Pentland [13] developed a Probabilistic Visual Learning (PVL) method which uses the eigenspace decomposition as an integral part of estimating the pdf in high-dimensional image spaces. To address the lack of sufficient training data Liu and Wechsler [12] introduced the Probabilistic Reasoning Models (PRM) where the conditional pdf for each class is modeled using the pooled within-class scatter and the Maximum A Posteriori (MAP) Bayes classification rule is implemented in the reduced PCA subspace.

The Fisher's Linear Discriminant (FLD) is yet another popular discriminant criterion. By applying first PCA for dimensionality reduction and then FLD for discriminant analysis, Belhumire, Hespanha, and Kriegman [2] and Swets and Weng [18] developed similar methods (Fisherfaces and the Most Discriminant Features (MDF) method) for face recognition. Methods that combine PCA and the standard FLD, however, lack in their generalization ability as they overfit to the training data. To address the overfitting problem Liu and Wechsler [11] introduced Enhanced FLD Models (EFM) to improve on the generalization capability of the standard FLD based classifiers such as Fisherfaces [2].

3. Independent Component Analysis (ICA)

As PCA considers the 2nd order moments only it lacks information on higher order statistics. ICA accounts for higher order statistics and it identifies the independent source components from their linear mixtures (the observables). ICA thus provides a more powerful data representation than PCA [9] as its goal is that of providing an independent rather than uncorrelated image decomposition and representation.

ICA of a random vector searches for a linear transformation which minimizes the statistical dependence between its components [5]. In particular, let $X \in \mathbb{R}^N$ be a random vector representing an image, where N is the dimensionality of the image space. The vector is formed by concatenating the rows or the columns of the image which may be normalized to have a unit norm and/or an equalized histogram. The covariance matrix of X is defined as

$$\Sigma_X = E\{[X - E(X)][X - E(X)]^t\} \tag{1}$$

where $E(\cdot)$ is the expectation operator, t denotes the transpose operation, and $\Sigma_X \in \mathbb{R}^{N \times N}$. The ICA of X factorizes the covariance matrix Σ_X into the following form

$$\Sigma_X = F \Delta F^t \tag{2}$$

where Δ is diagonal real positive and F transforms the original data X into Z

$$X = FZ \tag{3}$$

such that the components of the new data Z are independent or "the most independent possible" [5].

To derive the ICA transformation F, Comon [5] developed an algorithm which consists of three operations: whitening, rotation, and normalization. First, the whitening operation transforms a random vector X into another one U that has a unit covariance matrix.

$$X = \Phi \Lambda^{1/2} U \tag{4}$$

where Φ and Λ are derived by solving the following eigenvalue equation.

$$\Sigma_X = \Phi \Lambda \Phi^t \tag{5}$$

where $\Phi = [\phi_1, \phi_2, \ldots, \phi_N]$ is an orthonormal eigenvector matrix and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ is a diagonal eigenvalue matrix of Σ_X. One notes that whitening, an integral ICA component, counteracts the fact that the Mean Square Error (MSE) preferentially weighs low frequencies [14]. The rotation operations, then, perform source separation (to derive independent components) by minimizing the mutual information approximated using higher order cumulants. Finally, the normalization operation derives unique independent components in terms of orientation, unit norm, and order of projections [5].

4. Sensitivity Analysis of ICA

Eq. 4 can be rearranged to the following form

$$U = \Lambda^{-1/2}\Phi^t X \tag{6}$$

where Φ and Λ are eigenvector and eigenvalue matrices, respectively (see Eq. 5). Eq. 6 shows that during whitening the eigenvalues appear in the denominator. The trailing eigenvalues, which tend to capture noise as their values are fairly small, thus cause the whitening step to fit for misleading variations and make the ICA criterion to generalize poorly when it is exposed to new data. As a consequence, if the whitening step is preceded by a dimensionality reduction procedure ICA performance would be enhanced and computational complexity reduced. Specifically, space dimensionality is first reduced to discard the small trailing eigenvalues and only then the compressed data is normalized ('sphered') to unit gain control.

If we assume that the eigenvalues in Λ are sorted in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, then the first m ($m < N$) leading eigenvectors define a matrix $P \in \mathbb{R}^{N \times m}$

$$P = [\Phi_1, \Phi_2, \ldots, \Phi_m] \tag{7}$$

and the first m eigenvalues specify a diagonal matrix $\Lambda_1 \in \mathbb{R}^{m \times m}$

$$\Lambda_1 = diag\{\lambda_1, \lambda_2, \ldots, \lambda_m\} \tag{8}$$

The dimensionality reduction whitening transforms the data X into $V \in \mathbb{R}^m$

$$V = {\Lambda_1}^{-1/2} P^t X \tag{9}$$

The question now is how to choose the dimensionality m of the reduced subspace. Note that the goal of using whitening for dimensionality reduction is two-fold. On the one hand, we hope to lose as little representative information of the original data as possible in the transformation from the high dimensional space to the low dimensional one. On the other hand, in the reduced subspace the small trailing eigenvalues are excluded so that we can obtain more robust whitening results. Toward that end, during the whitening transformation we should keep a balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data and the requirement that the trailing eigenvalues of the covariance matrix are not too small. As a result, the eigenvalue spectrum of the training data supplies useful information for the decision of the dimensionality m of the subspace.

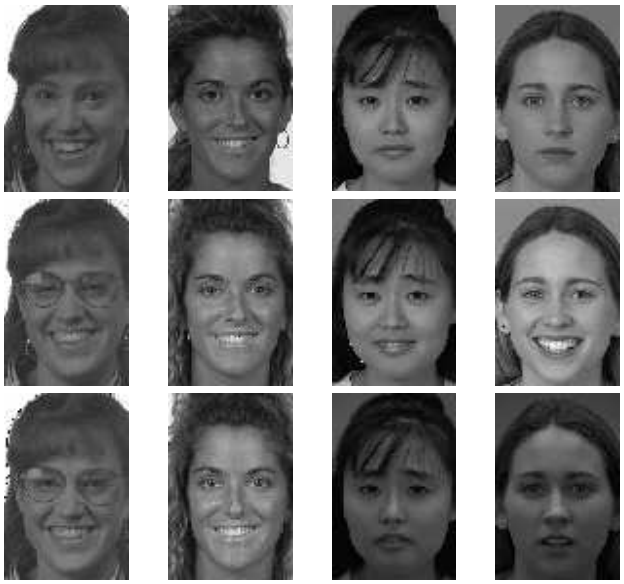

Figure 1. Some example images from the FERET database (cropped to the size of 64×96)

The experimental data, consisting of 1,107 facial images corresponding to 369 subjects, comes from the FERET database [16]. Some of the face images used are shown in Fig. 1. 600 out of the 1,107 images correspond to 200 subjects with each subject having three images — two of them are the first and the second shot, while the third shot is taken under low illumination. 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 and much later time. Two images of each subject are used for training with the remaining one used for testing. The images are cropped to the size of 64 x 96, once the eye coordinates are manually detected. Fig. 2 shows the relative magnitude of the eigenvalues derived using the 738 training face images. When the eigenvalue index is greater than 40, the corresponding eigenvalues have relatively small magnitudes, and if they were included in the whitening transformation these small eigenvalues will lead

to decreased ICA performance (see Fig. 3 and Fig. 4) as they amplify the effects of noise. So we set the dimension of the reduced space as $m = 40$.

In order to assess the sensitivity of ICA in terms of the dimension of the compressed and whitened space where it is implemented, we carried out a comparative assessment for different dimensional (m) whitened subspace. Fig. 3 and Fig. 4 show the ICA face recognition performance when different m is chosen during the whitening transformation (see Eq. 9) and as the number of features used can range up to the dimension of the compressed space. The curve corresponding to $m = 40$ performs better than all the other curves obtained for different m values. Fig. 4 shows that ICA performance deteriorates quite severely as m is getting larger. The reason for this deterioration is that for the large values of m, more small trailing eigenvalues (see Fig. 2) are included in the whitening step and this leads to an unstable transformation. Note that smaller values for m lead to decreased ICA performance as well because the whitened space fails to capture enough information on the original data.

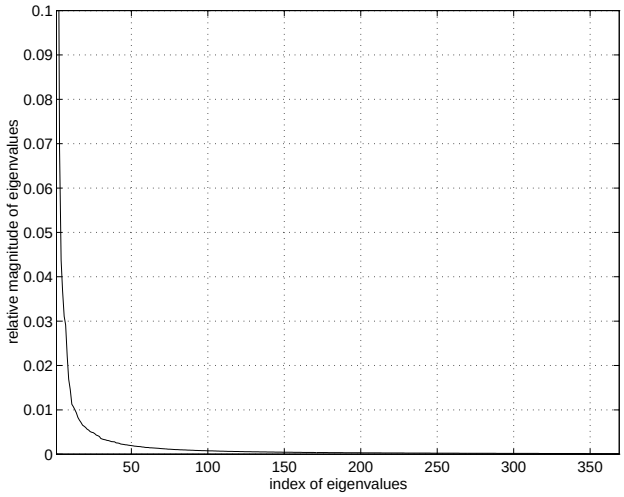

Figure 2. The relative magnitude $(\lambda_i / \sum_k \lambda_k)$ of eigenvalues

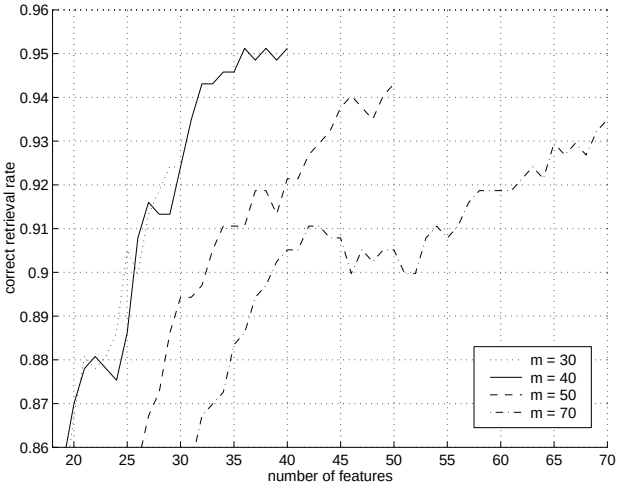

Figure 3. Comparative performance of ICA as a discriminant criterion when $m = 30, 40, 50, 70$, respectively.

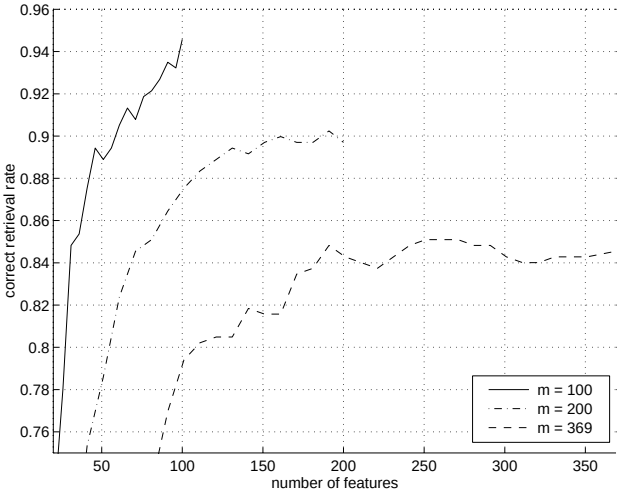

Figure 4. Comparative performance of ICA as a discriminant criterion when $m = 100, 200, 369$, respectively.

5. Discrimination Power of ICA

To assess the performance of ICA as a discriminant criterion, we implemented the eigenfaces [19] and Fisherfaces [2] methods for comparison purposes. Note that Fisherfaces apply first PCA for dimensionality reduction and then FLD for discriminant analysis. Relevant questions concerning PCA are usually related to the range of principal components used and how it affects performance. Regarding discriminant analysis one has to understand the reasons for overfitting and how to avoid it. One can actually show that using more principal components may lead to decreased performance (for recognition). The explanation for this behavior is that the trailing eigenvalues correspond to high-frequency components and usually encode noise. As a result, when these trailing but small valued eigenvalues are used to define the reduced PCA subspace, the FLD procedure has to fit for noise as well and as a

consequence overfitting takes place. To improve the generalization ability of Fisherfaces, we implemented it in the $m = 200$ PCA subspace. The comparative performance of eigenfaces, Fisherfaces and ICA is plotted in Fig. 5 when ICA was implemented in the $m = 40$ PCA subspace. Fig. 5 shows that ICA criterion performs better than both eigenfaces and Fisherfaces.

We also assessed the ICA discriminant criterion against two other popular discriminant criteria: the MAP rule of the Bayes classifier and the Fisher's linear discriminant as they are embedded within the Probabilistic Reasoning Models (PRM) [12] and the Enhanced FLD Models (EFM) [11]. Fig. 6 plots the comparative performances of PRM-1 and EFM-1 against the ICA method with m again being set to 40. ICA is shown to have comparable face recognition performance with the MAP rule of the Bayes classifier and the Fisher's linear discriminant as embedded within PRM and EFM, respectively.

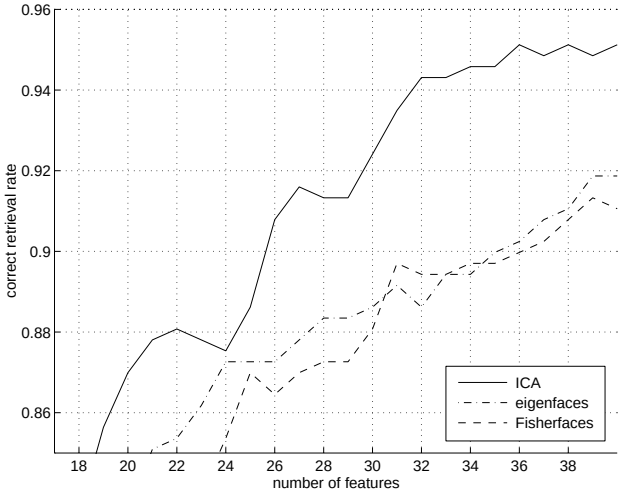

Figure 5. Comparative performance of eigenfaces, Fisherfaces, and ICA.

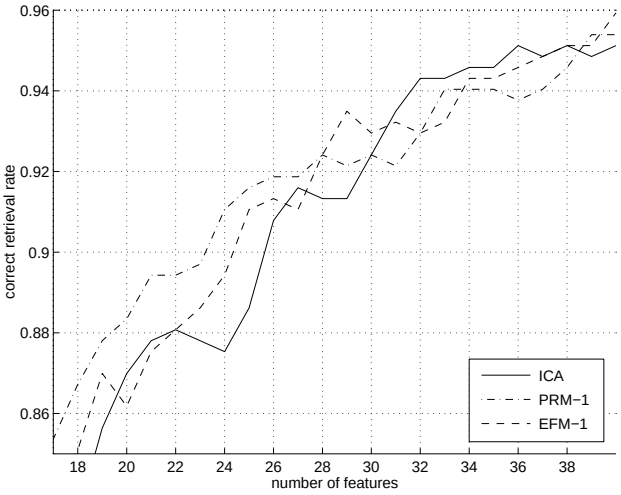

Figure 6. Comparative performance of PRM-1, EFM-1, and ICA.

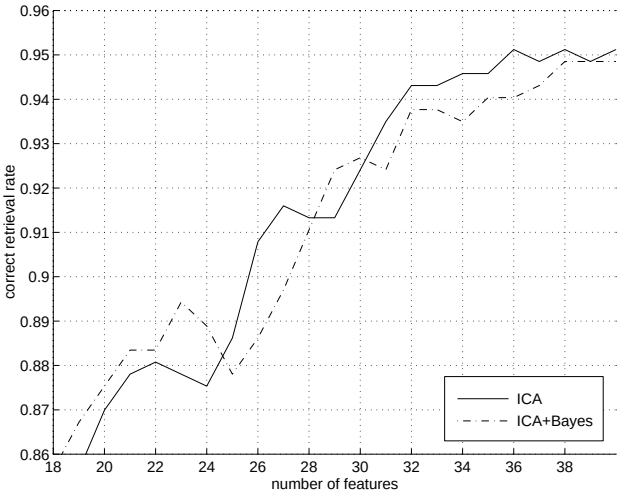

Figure 7. Comparative performance of ICA and ICA combined with Bayes classifier.

We next augmented the ICA criterion by additional criteria such as the MAP rule of the Bayes classifier or the FLD. In the ICA space, the Bayes classifier uses the pooled within-class scatter to estimate the covariance matrix for each class in order to approximate the conditional pdf, and then applies the MAP rule as the classification criterion (see [12] for detail). The FLD also use the pooled within-class scatter to estimate the within-class covariance matrix in the ICA space [11]. The first augmented criterion (ICA + Bayes classifier) does not improve the face recognition rate as it displays a similar performance curve to that of ICA as plotted in Fig. 7. Note that when ICA is combined with the Bayes classifier, the MAP rule specifies a quadratic classifier characterized by Mahalanobis distance. The whitening transformation to standardize the data is applied first and it precedes MAP. As a result, the Mahalanobis distance embedded in the MAP classifier duplicates to some extent the whitening component of ICA and can not improve the overall performance. The second augmented criterion (ICA + FLD), whose performance is shown in Fig. 8, significantly deteriorates the recognition performance. This deterioration is caused by the additional FLD transformation which cancels to a large extent the independence criterion intrinsic to ICA.

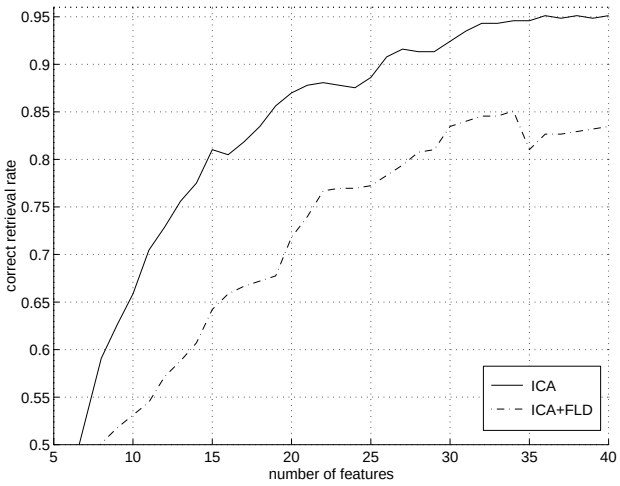

Figure 8. Comparative performance of ICA and ICA combined with FLD.

6. Conclusions

This paper addresses the relative usefulness of the independent component analysis for Face Recognition. Comparative assessments are made regarding (i) ICA sensitivity to the dimension of the space where it is carried out, and (ii) ICA discriminant performance alone or when combined with other discriminant criteria such as the MAP criteria of the Bayes classifier or the Fisher's linear discriminant. The sensitivity analysis suggests that for enhanced performance ICA should be carried out in a compressed and whitened space where most of the representative information of the original data is preserved and the small trailing eigenvalues discarded. The dimensionality of the compressed subspace is decided based on the eigenvalue spectrum from the training data. The discriminant analysis shows that the ICA criterion, when carried out in the properly compressed and whitened space, performs better than the eigenfaces and Fisherfaces methods for face recognition, but its performance deteriorates significantly when augmented by an additional discriminant criteria such as the FLD.

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] M. Bartlett and T. Sejnowski. Independent components of face images: A representation for face recognition. In *Proc. the 4th Annual Joint Symposium on Neural Computation*, Pasadena, CA, May 17, 1997.

[2] 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.

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

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

[5] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

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

[7] A. Hyvarinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9:1483–1492, 1997.

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

[9] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo. A class of neural networks for independent component analysis. *IEEE Trans. on Neural Networks*, 8(3):486–504, 1997.

[10] M. Kirby and L. Sirovich. Application of the karhunenloeve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[11] C. Liu and H. Wechsler. Enhanced fisher linear discriminant models for face recognition. In *Proc. the 14th International Conference on Pattern Recognition*, Queensland, Australia, August 17-20, 1998.

[12] C. Liu and H. Wechsler. Probabilistic reasoning models for face recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, USA, June 23-25, 1998.

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

[14] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(13):607–609, 1996.

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

[16] 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.

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

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

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