Face Recognition by Independent Component Analysis

Marian Stewart Bartlett, *Member, IEEE*, Javier R. Movellan, *Member, IEEE*, and Terrence J. Sejnowski, *Fellow, IEEE*

***Abstract*—A number of current face recognition algorithms use face representations found by unsupervised statistical methods. Typically these methods find a set of basis images and represent faces as a linear combination of those images. Principal component analysis (PCA) is a popular example of such methods. The basis images found by PCA depend only on pairwise relationships between pixels in the image database. In a task such as face recognition, in which important information may be contained in the high-order relationships among pixels, it seems reasonable to expect that better basis images may be found by methods sensitive to these high-order statistics. Independent component analysis (ICA), a generalization of PCA, is one such method. We used a version of ICA derived from the principle of optimal information transfer through sigmoidal neurons. ICA was performed on face images in the FERET database under two different architectures, one which treated the images as random variables and the pixels as outcomes, and a second which treated the pixels as random variables and the images as outcomes. The first architecture found spatially local basis images for the faces. The second architecture produced a factorial face code. Both ICA representations were superior to representations based on PCA for recognizing faces across days and changes in expression. A classifier that combined the two ICA representations gave the best performance.**

***Index Terms*—Eigenfaces, face recognition, independent component analysis (ICA), principal component analysis (PCA), unsupervised learning.**

I. Introduction

REDUNDANCY in the sensory input contains structural information about the environment. Barlow has argued that such redundancy provides knowledge [5] and that the role of the sensory system is to develop factorial representations in which these dependencies are separated into independent components (ICs). Barlow also argued that such representations are advantageous for encoding complex objects that are characterized by high-order dependencies. Atick and Redlich have also argued for such representations as a general coding strategy for the visual system [3].

Principal component analysis (PCA) is a popular unsupervised statistical method to find useful image representations. Consider a set of n basis images each of which has n pixels. A standard basis set consists of a single active pixel with intensity 1, where each basis image has a different active pixel. Any given image with n pixels can be decomposed as a linear combination of the standard basis images. In fact, the pixel values of an image can then be seen as the coordinates of that image with respect to the standard basis. The goal in PCA is to find a "better" set of basis images so that in this new basis, the image coordinates (the PCA coefficients) are uncorrelated, i.e., they cannot be linearly predicted from each other. PCA can, thus, be seen as partially implementing Barlow's ideas: Dependencies that show up in the joint distribution of pixels are separated out into the marginal distributions of PCA coefficients. However, PCA can only separate pairwise linear dependencies between pixels. High-order dependencies will still show in the joint distribution of PCA coefficients, and, thus, will not be properly separated.

Some of the most successful representations for face recognition, such as eigenfaces [57], holons [15], and local feature analysis [50] are based on PCA. In a task such as face recognition, much of the important information may be contained in the high-order relationships among the image pixels, and thus, it is important to investigate whether generalizations of PCA which are sensitive to high-order relationships, not just second-order relationships, are advantageous. Independent component analysis (ICA) [14] is one such generalization. A number of algorithms for performing ICA have been proposed. See [20] and [29] for reviews. Here, we employ an algorithm developed by Bell and Sejnowski [11], [12] from the point of view of optimal information transfer in neural networks with sigmoidal transfer functions. This algorithm has proven successful for separating randomly mixed auditory signals (the cocktail party problem), and for separating electroencephalogram (EEG) signals [37] and functional magnetic resonance imaging (fMRI) signals [39].

We performed ICA on the image set under two architectures. Architecture I treated the images as random variables and the pixels as outcomes, whereas Architecture II treated the

Manuscript received May 21, 2001; revised May 8, 2002. This work was supported by University of California Digital Media Innovation Program D00-10084, the National Science Foundation under Grants 0086107 and IIT-0223052, the National Research Service Award MH-12417-02, the Lawrence Livermore National Laboratories ISCR agreement B291528, and the Howard Hughes Medical Institute. An abbreviated version of this paper appears in *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III, Vol. 3299*, B. Rogowitz and T. Pappas, Eds., 1998. Portions of this paper use the FERET database of facial images, collected under the FERET program of the Army Research Laboratory.

The authors are with the University of California-San Diego, La Jolla, CA 92093-0523 USA (e-mail: marni@salk.edu; javier@inc.ucsd.edu; terry@salk.edu).

T. J. Sejnowski is also with the Howard Hughes Medical Institute at the Salk Institute, La Jolla, CA 92037 USA.

Digital Object Identifier 10.1109/TNN.2002.804287

pixels as random variables and the images as outcomes.[1] Matlab code for the ICA representations is available at http://inc.ucsd.edu/~marni.

Face recognition performance was tested using the FERET database [52]. Face recognition performances using the ICA representations were benchmarked by comparing them to performances using PCA, which is equivalent to the "eigenfaces" representation [51], [57]. The two ICA representations were then combined in a single classifier.

II. ICA

There are a number of algorithms for performing ICA [11], [13], [14], [25]. We chose the infomax algorithm proposed by Bell and Sejnowski [11], which was derived from the principle of optimal information transfer in neurons with sigmoidal transfer functions [27]. The algorithm is motivated as follows: Let $\mathbf{X}$ be an n-dimensional (n-D) random vector representing a distribution of inputs in the environment. (Here, boldface capitals denote random variables, whereas plain text capitals denote matrices). Let W be an $n \times n$ invertible matrix, $\mathbf{U} = W\mathbf{X}$ and $\mathbf{Y} = f(\mathbf{U})$ an n-D random variable representing the outputs of n-neurons. Each component of $f = (f_1, \ldots, f_n)$ is an invertible squashing function, mapping real numbers into the $[0, 1]$ interval. Typically, the logistic function is used

$$f_i(u) = \frac{1}{1 + e^{-u}}. \quad (1)$$

The $\mathbf{U}_1, \ldots, \mathbf{U}_n$ variables are linear combinations of inputs and can be interpreted as presynaptic activations of n-neurons. The $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ variables can be interpreted as postsynaptic activation rates and are bounded by the interval $[0, 1]$. The goal in Bell and Sejnowski's algorithm is to maximize the mutual information between the environment $\mathbf{X}$ and the output of the neural network $\mathbf{Y}$. This is achieved by performing gradient ascent on the entropy of the output with respect to the weight matrix W. The gradient update rule for the weight matrix, W is as follows:

$$\Delta W \propto \nabla_W H(\mathbf{Y}) = (W^T)^{-1} + E(\mathbf{Y}'\mathbf{X}^T) \quad (2)$$

where $\mathbf{Y}'_i = f''_i(\mathbf{U}_i)/f'_i(\mathbf{U}_i)$, the ratio between the second and first partial derivatives of the activation function, T stands for transpose, E for expected value, $H(\mathbf{Y})$ is the entropy of the random vector $\mathbf{Y}$, and $\nabla_W H(\mathbf{Y})$ is the gradient of the entropy in matrix form, i.e., the cell in row i, column j of this matrix is the derivative of $H(\mathbf{Y})$ with respect to W_{ij}. Computation of the matrix inverse can be avoided by employing the natural gradient [1], which amounts to multiplying the absolute gradient by $W^T W$, resulting in the following learning rule [12]:

$$\Delta W \propto \nabla_W H(\mathbf{Y}) W^T W = (I + \mathbf{Y}'\mathbf{U}^T) W \quad (3)$$

where I is the identity matrix. The logistic transfer function (1) gives $\mathbf{Y}'_i = (1 - 2\mathbf{Y}_i)$.

When there are multiple inputs and outputs, maximizing the joint entropy of the output $\mathbf{Y}$ encourages the individual outputs to move toward statistical independence. When the form of the nonlinear transfer function f is the same as the cumulative density functions of the underlying ICs (up to scaling and translation) it can be shown that maximizing the joint entropy of the outputs in $\mathbf{Y}$ also minimizes the mutual information between the individual outputs in $\mathbf{U}$ [12], [42]. In practice, the logistic transfer function has been found sufficient to separate mixtures of natural signals with sparse distributions including sound sources [11].

The algorithm is speeded up by including a "sphering" step prior to learning [12]. The row means of $\mathbf{X}$ are subtracted, and then $\mathbf{X}$ is passed through the whitening matrix W_z, which is twice the inverse square root[2] of the covariance matrix

$$W_z = 2 * (Cov(\mathbf{X}))^{-(1/2)}. \quad (4)$$

This removes the first and the second-order statistics of the data; both the mean and covariances are set to zero and the variances are equalized. When the inputs to ICA are the "sphered" data, the full transform matrix W_I is the product of the sphering matrix and the matrix learned by ICA

$$W_I = W W_z. \quad (5)$$

MacKay [36] and Pearlmutter [48] showed that the ICA algorithm converges to the maximum likelihood estimate of W^{-1} for the following generative model of the data:

$$\mathbf{X} = W^{-1}\mathbf{S} \quad (6)$$

where $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_n)'$ is a vector of independent random variables, called the sources, with cumulative distributions equal to f_i, in other words, using logistic activation functions corresponds to assuming logistic random sources and using the standard cumulative Gaussian distribution as activation functions corresponds to assuming Gaussian random sources. Thus, W^{-1}, the inverse of the weight matrix in Bell and Sejnowski's algorithm, can be interpreted as the source mixing matrix and the $\mathbf{U} = W\mathbf{X}$ variables can be interpreted as the maximum-likelihood (ML) estimates of the sources that generated the data.

A. ICA and Other Statistical Techniques

ICA and PCA: PCA can be derived as a special case of ICA which uses Gaussian source models. In such case the mixing matrix $\boldsymbol{W}$ is unidentifiable in the sense that there is an infinite number of equally good ML solutions. Among all possible ML solutions, PCA chooses an orthogonal matrix which is optimal in the following sense: 1) Regardless of the distribution of $\mathbf{X}$, $\mathbf{U}_1$ is the linear combination of input that allows optimal linear reconstruction of the input in the mean square sense; and 2) for $\mathbf{U}_1, \ldots \mathbf{U}_k$ fixed, $\mathbf{U}_{k+1}$ allows optimal linear reconstruction among the class of linear combinations of $\mathbf{X}$ which are uncorrelated with $\mathbf{U}_1 \cdots \mathbf{U}_k$. If the sources are Gaussian, the likelihood of the data depends only on first- and second-order statistics (the covariance matrix). In PCA, the rows of W are, in fact, the eigenvectors of the covariance matrix of the data.

Second-order statistics capture the amplitude spectrum of images but not their phase spectrum. The high-order statistics capture the phase spectrum [12], [19]. For a given sample

[1]Preliminary versions of this work appear in [7] and [9]. A longer discussion of unsupervised learning for face recognition appears in [6].

[2]We use the principal square root, which is the unique square root for which every eigenvalue has nonnegative real part.

of natural images, we can scramble their phase spectrum while maintaining their power spectrum. This will dramatically alter the appearance of the images but will not change their second-order statistics. The phase spectrum, not the power spectrum, contains the structural information in images that drives human perception. For example, as illustrated in Fig. 1, a face image synthesized from the amplitude spectrum of face A and the phase spectrum of face B will be perceived as an image of face B [45], [53]. The fact that PCA is only sensitive to the power spectrum of images suggests that it might not be particularly well suited for representing natural images.

The assumption of Gaussian sources implicit in PCA makes it inadequate when the true sources are non-Gaussian. In particular, it has been empirically observed that many natural signals, including speech, natural images, and EEG are better described as linear combinations of sources with long tailed distributions [11], [19]. These sources are called "high-kurtosis," "sparse," or "super-Gaussian" sources. Logistic random variables are a special case of sparse source models. When sparse source models are appropriate, ICA has the following potential advantages over PCA: 1) It provides a better probabilistic model of the data, which better identifies where the data concentrate in n-dimensional space. 2) It uniquely identifies the mixing matrix W. 3) It finds a not-necessarily orthogonal basis which may reconstruct the data better than PCA in the presence of noise. 4) It is sensitive to high-order statistics in the data, not just the covariance matrix.

Fig. 2 illustrates these points with an example. The figure shows samples from a three-dimensional (3-D) distribution constructed by linearly mixing two high-kurtosis sources. The figure shows the basis vectors found by PCA and by ICA on this problem. Since the three ICA basis vectors are nonorthogonal, they change the relative distance between data points. This change in metric may be potentially useful for classification algorithms, like nearest neighbor, that make decisions based on relative distances between points. The ICA basis also alters the angles between data points, which affects similarity measures such as cosines. Moreover, if an undercomplete basis set is chosen, PCA and ICA may span different subspaces. For example, in Fig. 2, when only two dimensions are selected, PCA and ICA choose different subspaces.

The metric induced by ICA is superior to PCA in the sense that it may provide a representation more robust to the effect of noise [42]. It is, therefore, possible for ICA to be better than PCA for reconstruction in noisy or limited precision environments. For example, in the problem presented in Fig. 2, we found that if only 12 bits are allowed to represent the PCA and ICA coefficients, linear reconstructions based on ICA are 3 dB better than reconstructions based on PCA (the noise power is reduced by more than half). A similar result was obtained for PCA and ICA subspaces. If only four bits are allowed to represent the first 2 PCA and ICA coefficients, ICA reconstructions are 3 dB better than PCA reconstructions. In some problems, one can think of the actual inputs as noisy versions of some canonical inputs. For example, variations in lighting and expressions can be seen as noisy versions of the canonical image of a person. Having input representations which are robust to noise may potentially give us representations that better reflect the data.

Fig. 1. (left) Two face images. (Center) The two faces with scrambled phase. (right) Reconstructions with the amplitude of the original face and the phase of the other face. Faces images are from the FERET face database, reprinted with permission from J. Phillips.

When the sources models are sparse, ICA is closely related to the so called nonorthogonal "rotation" methods in PCA and factor analysis. The goal of these rotation methods is to find directions with high concentrations of data, something very similar to what ICA does when the sources are sparse. In such cases, ICA can be seen as a theoretically sound probabilistic method to find interesting nonorthogonal "rotations."

ICA and Cluster Analysis: Cluster analysis is a technique for finding regions in n-dimensional space with large concentrations of data. These regions are called "clusters." Typically, the main statistic of interest in cluster analysis is the center of those clusters. When the source models are sparse, ICA finds directions along which significant concentrations of data points are observed. Thus, when using sparse sources, ICA can be seen as a form of cluster analysis. However, the emphasis in ICA is on finding optimal directions, rather than specific locations of high data density. Fig. 2 illustrates this point. Note how the data concentrates along the ICA solutions, not the PCA solutions. Note also that in this case, all the clusters have equal mean, and thus are better characterized by their orientation rather than their position in space.

It should be noted that ICA is a very general technique. When super-Gaussian sources are used, ICA can be seen as doing something akin to nonorthogonal PCA and to cluster analysis, however, when the source models are sub-Gaussian, the relationship between these techniques is less clear. See [30] for a discussion of ICA in the context of sub-Gaussian sources.

B. Two Architectures for Performing ICA on Images

Let X be a data matrix with n_r rows and n_c columns. We can think of each column of X as outcomes (independent trials) of a random experiment. We think of the ith row of X as the specific value taken by a random variable $\mathbf{X}_i$ across n_c independent trials. This defines an empirical probability distribution for $\mathbf{X}_1, \ldots \mathbf{X}_{n_r}$ in which each column of X is given probability mass $1/n_c$. Independence is then defined with respect to such

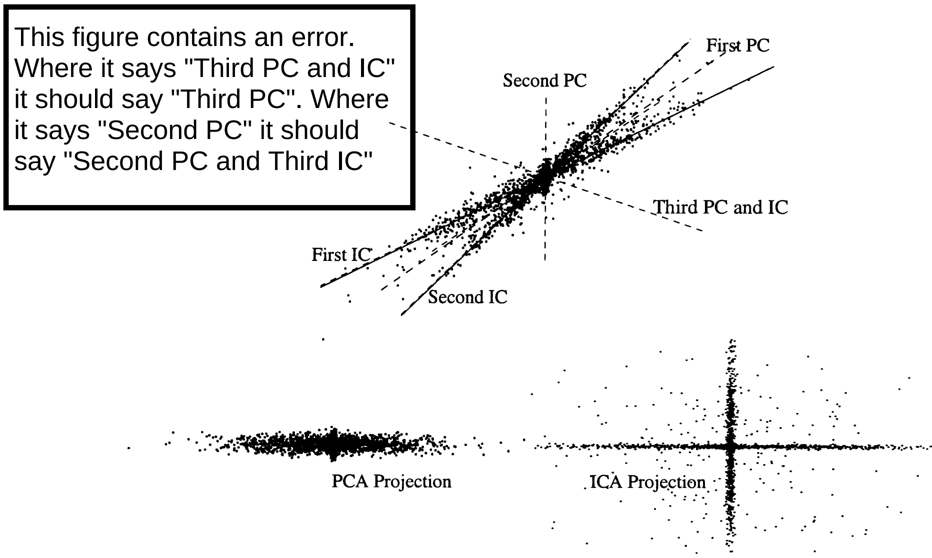


Fig. 2. (top) Example 3-D data distribution and corresponding PC and IC axes. Each axis is a column of the mixing matrix W^{-1} found by PCA or ICA. Note the PC axes are orthogonal while the IC axes are not. If only two components are allowed, ICA chooses a different subspace than PCA. (bottom left) Distribution of the first PCA coordinates of the data. (bottom right) Distribution of the first ICA coordinates of the data. Note that since the ICA axes are nonorthogonal, relative distances between points are different in PCA than in ICA, as are the angles between points.

a distribution. For example, we say that rows i and j of X are independent if it is not possible to predict the values taken by $\mathbf{X}_j$ across columns from the corresponding values taken by $\mathbf{X}_i$, i.e.,

$$P(\mathbf{X}_i = u, \mathbf{X}_j = v) = P(\mathbf{X}_i = u)P(\mathbf{X}_j = v) \qquad \text{for all } u, v \in R \tag{7}$$

where P is the empirical distribution as in (7).

Our goal in this paper is to find a good set of basis images to represent a database of faces. We organize each image in the database as a long vector with as many dimensions as number of pixels in the image. There are at least two ways in which ICA can be applied to this problem.

1) We can organize our database into a matrix X where each row vector is a different image. This approach is illustrated in (Fig. 3 left). In this approach, images are random variables and pixels are trials. In this approach, it makes sense to talk about independence of images or functions of images. Two images i and j are independent if when moving across pixels, it is not possible to predict the value taken by the pixel on image j based on the value taken by the same pixel on image i. A similar approach was used by Bell and Sejnowski for sound source separation [11], for EEG analysis [37], and for fMRI [39].
2) We can transpose X and organize our data so that images are in the columns of X. This approach is illustrated in (Fig. 3 right). In this approach, pixels are random variables and images are trials. Here, it makes sense to talk about independence of pixels or functions of pixels. For example, pixel i and j would be independent if when moving across the entire set of images it is not possible to predict the value taken by pixel i based on the corresponding value taken by pixel j on the same image. This approach was inspired by Bell and Sejnowski's work on the ICs of natural images [12].

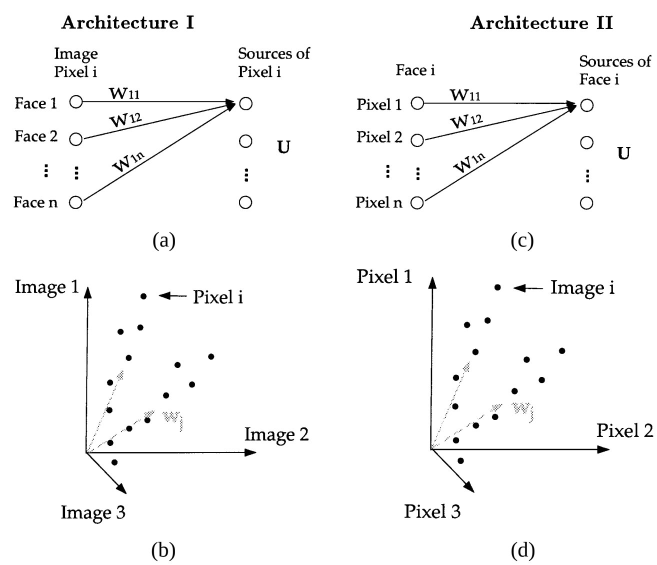


Fig. 3. Two architectures for performing ICA on images. (a) Architecture I for finding statistically independent basis images. Performing source separation on the face images produced IC images in the rows of U. (b) The gray values at pixel location i are plotted for each face image. ICA in architecture I finds weight vectors in the directions of statistical dependencies among the pixel locations. (c) Architecture II for finding a factorial code. Performing source separation on the pixels produced a factorial code in the columns of the output matrix, U. (d) Each face image is plotted according to the gray values taken on at each pixel location. ICA in architecture II finds weight vectors in the directions of statistical dependencies among the face images.

III. Image Data

The face images employed for this research were a subset of the FERET face database [52]. The data set contained images of 425 individuals. There were up to four frontal views of each individual: A neutral expression and a change of expression from one session, and a neutral expression and change of expression from a second session that occurred up to two years after the first. Examples of the four views are shown in Fig. 6. The algorithms were trained on a single frontal view of each

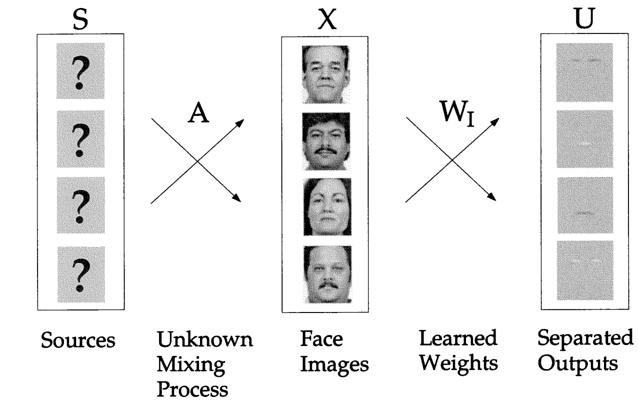


Fig. 4. Image synthesis model for Architecture I. To find a set of IC images, the images in X are considered to be a linear combination of statistically independent basis images, S, where A is an unknown mixing matrix. The basis images were estimated as the learned ICA output U.

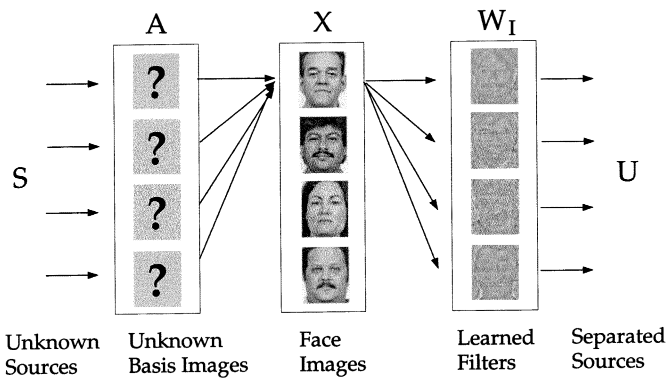


Fig. 5. Image synthesis model for Architecture II, based on [43] and [44]. Each image in the dataset was considered to be a linear combination of underlying basis images in the matrix A. The basis images were each associated with a set of independent "causes," given by a vector of coefficients in S. The basis images were estimated by $\mathbf{A} = \mathbf{W}_I^{-1}$, where $\mathbf{W}_I$ is the learned ICA weight matrix.

individual. The training set was comprised of 50% neutral expression images and 50% change of expression images. The algorithms were tested for recognition under three different conditions: same session, different expression; different day, same expression; and different day, different expression (see Table I).

Coordinates for eye and mouth locations were provided with the FERET database. These coordinates were used to center the face images, and then crop and scale them to 60 $\times$ 50 pixels. Scaling was based on the area of the triangle defined by the eyes and mouth. The luminance was normalized by linearly rescaling each image to the interval [0, 255]. For the subsequent analyses, each image was represented as a 3000–dimensional vector given by the luminance value at each pixel location.

IV. Architecture I: Statistically Independent Basis Images

As described earlier, the goal in this approach is to find a set of statistically independent basis images. We organize the data matrix X so that the images are in rows and the pixels are in columns, i.e., X has 425 rows and 3000 columns, and each image has zero mean.

Fig. 6. Example from the FERET database of the four frontal image viewing conditions: neutral expression and change of expression from session 1; neutral expression and change of expression from session 2. Reprinted with permission from Jonathan Phillips.

TABLE I
Image Sets Used for Training and Testing

Image Set	Condition		No. Images
Training Set	Session I	50% neutral 50% other	425
Test Set 1	Same Day	Different Expression	421
Test Set 2	Different Day	Same Expression	45
Test Set 3	Different Day	Different Expression	43

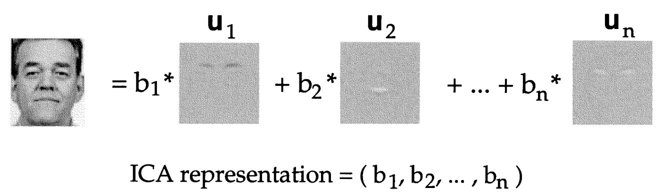


Fig. 7. The independent basis image representation consisted of the coefficients, $\mathbf{b}$, for the linear combination of independent basis images, $\mathbf{u}$, that comprised each face image $\mathbf{x}$.

In this approach, ICA finds a matrix W such that the rows of $U = WX$ are as statistically independent as possible. The source images estimated by the rows of U are then used as basis images to represent faces. Face image representations consist of the coordinates of these images with respect to the image basis defined by the rows of U, as shown in Fig. 7. These coordinates are contained in the mixing matrix $A \triangleq W_I^{-1}$.

The number of ICs found by the ICA algorithm corresponds to the dimensionality of the input. Since we had 425 images in the training set, the algorithm would attempt to separate 425 ICs. Although we found in previous work that performance improved with the number of components separated, 425 was intractable under our present memory limitations. In order to have control over the number of ICs extracted by the algorithm, instead of performing ICA on the n_r original images, we performed ICA on a set of m linear combinations of those images, where $m < n_r$. Recall that the image synthesis model assumes that the images in X are a linear combination of a set of unknown statistically independent sources. The image synthesis model is unaffected by replacing the original images with some other linear combination of the images.

Adopting a method that has been applied to ICA of fMRI data [39], we chose for these linear combinations the first m PC eigenvectors of the image set. PCA on the image set in which the pixel locations are treated as observations and each face image a measure, gives the linear combination of the parameters (images) that accounts for the maximum variability in the observations (pixels). The use of PCA vectors in the input did not throw away the high-order relationships. These relationships still existed in the data but were not separated.

Let P_m denote the matrix containing the first m PC axes in its columns. We performed ICA on P_m^T, producing a matrix of m independent source images in the rows of U. In this implementation, the coefficients $\mathbf{b}$ for the linear combination of basis images in U that comprised the face images in X were determined as follows.

The PC representation of the set of zero-mean images in X based on P_m is defined as $R_m = XP_m$. A minimum squared error approximation of X is obtained by $\hat{X} = R_m P_m^T$.

The ICA algorithm produced a matrix $W_I = WW_Z$ such that

$$\begin{aligned} W_I P_m^T &= U \\ P_m^T &= W_I^{-1} U. \end{aligned} \tag{8}$$

Therefore

$$\begin{aligned} \hat{X} &= R_m P_m^T \\ \hat{X} &= R_m W_I^{-1} U. \end{aligned} \tag{9}$$

where W_Z was the sphering matrix defined in (4). Hence, the rows of $R_m\, W_I^{-1}$ contained the coefficients for the linear combination of statistically independent sources U that comprised $\hat{X}$, where $\hat{X}$ was a minimum squared error approximation of X, just as in PCA. The IC representation of the face images based on the set of m statistically independent feature images, U was, therefore, given by the rows of the matrix

$$B = R_m W_I^{-1}. \tag{10}$$

A representation for test images was obtained by using the PC representation based on the training images to obtain $R_{\text{test}} = X_{\text{test}} P_m$, and then computing

$$B_{\text{test}} = R_{\text{test}} W_I^{-1}. \tag{11}$$

Note that the PCA step is not required for the ICA representation of faces. It was employed to serve two purposes: 1) to reduce the number of sources to a tractable number and 2) to provide a convenient method for calculating representations of test images. Without the PCA step, $B = W_I^{-1}$ and $B_{\text{test}} = X_{\text{test}}(U)^{\dagger}$.[3]

The PC axes of the training set were found by calculating the eigenvectors of the pixelwise covariance matrix over the set of face images. ICA was then performed on the first $m = 200$ of these eigenvectors, where the first 200 PCs accounted for over 98% of the variance in the images.[4] The 1×3000 eigenvectors in P_{200} comprised the rows of the 200×3000 input matrix X. The input matrix X was sphered[5] according to (4), and the weights W were updated according to (3) for 1900 iterations. The learning rate was initialized at 0.0005 and annealed down to 0.0001. Training took 90 minutes on a Dec Alpha 2100a. Following training, a set of statistically independent source images were contained in the rows of the output matrix U.

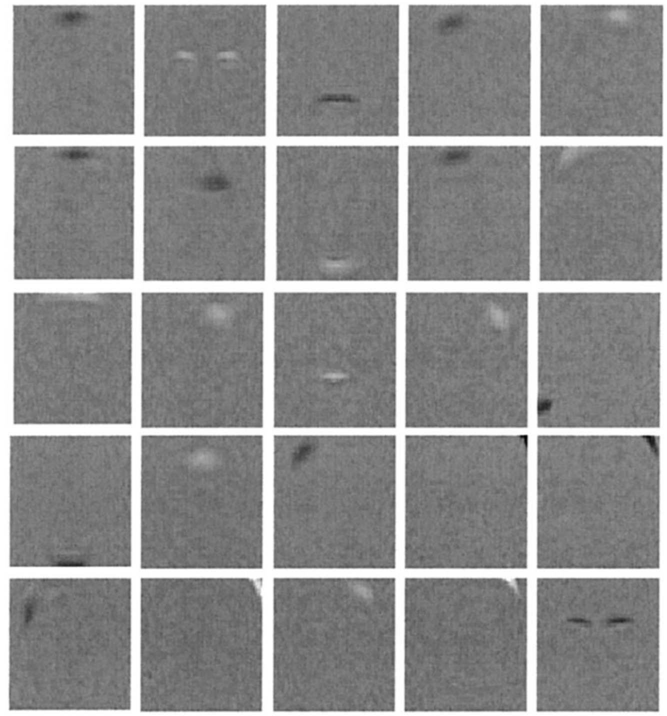


Fig. 8. Twenty-five ICs of the image set obtained by Architecture I, which provide a set of statistically independent basis images (rows of U in Fig. 4). ICs are ordered by the class discriminability ratio, r (4).

Fig. 8 shows a subset of 25 basis images (i.e., rows of U). These images can be interpreted as follows: Each row of the mixing matrix W found by ICA represents a cluster of pixels that have similar behavior across images. Each row of the U matrix tell us how close each pixel is to the cluster i identified by ICA. Since we use a sparse independent source model, these basis images are expected to be sparse and independent. Sparseness in this case means that the basis images will have a large number of pixels close to zero and a few pixels with large positive or negative values. Note that the ICA images are also local (regions with nonzero pixels are nearby). This is because a majority of the statistical dependencies are in spatially proximal pixel locations. A set of PC basis images (PCA axes) are shown in Fig. 9 for comparison.

A. Face Recognition Performance: Architecture I

Face recognition performance was evaluated for the coefficient vectors $\mathbf{b}$ by the nearest neighbor algorithm, using cosines as the similarity measure. Coefficient vectors in each test set were assigned the class label of the coefficient vector in the training set that was most similar as evaluated by the cosine of the angle between them

$$c = \frac{\mathbf{b}_{\text{test}} \cdot \mathbf{b}_{\text{train}}}{\|\mathbf{b}_{\text{test}}\| \|\mathbf{b}_{\text{train}}\|}. \tag{12}$$

Face recognition performance for the PC representation was evaluated by an identical procedure, using the PC coefficients contained in the rows of R_{200}.

[3] B_{test} could potentially be obtained without calculating a pseudoinverse by normalizing the length of the rows of U, thereby making U approximately orthonormal, and calculating $B_{\text{test}} = X_{\text{test}}\, U^T$. However, if ICA did not remove all of the second-order dependencies then U will not be precisely orthonormal.

[4] In pilot work, we found that face recognition performance improved with the number of components separated. We chose 200 components as the largest number to separate within our processing limitations.

[5] Although PCA already removed the covariances in the data, the variances were not equalized. We, therefore, retained the sphering step.

Fig. 9. First 25 PC axes of the image set (columns of P), ordered left to right, top to bottom, by the magnitude of the corresponding eigenvalue.

In experiments to date, ICA performs significantly better using cosines rather than Euclidean distance as the similarity measure, whereas PCA performs the same for both. A cosine similarity measure is equivalent to length-normalizing the vectors prior to measuring Euclidean distance when doing nearest neighbor

$$\begin{aligned} d^2(x,y) =& \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}\cdot\mathbf{y} \\ =& \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\|\cos(\alpha). \\ \text{Thus, if } & \|\mathbf{x}\| = \|\mathbf{y}\| = 1 \\ \min_{\mathbf{y}} d^2(x,y) =& \max_{\mathbf{y}} \cos(\alpha). \end{aligned} \tag{13}$$

Such contrast normalization is consistent with neural models of primary visual cortex [23]. Cosine similarity measures were previously found to be effective for computational models of language [24] and face processing [46].

Fig. 10 gives face recognition performance with both the ICA and the PCA-based representations. Recognition performance is also shown for the PCA-based representation using the first 20 PC vectors, which was the eigenface representation used by Pentland *et al.* [51]. Best performance for PCA was obtained using 200 coefficients. Excluding the first one, two, or three PCs did not improve PCA performance, nor did selecting intermediate ranges of components from 20 through 200. There was a trend for the ICA representation to give superior face recognition performance to the PCA representation with 200 components. The difference in performance was statistically significant for Test Set 3 ($Z = 1.94$, $p = 0.05$). The difference in performance between the ICA representation and the eigenface representation with 20 components was statistically significant

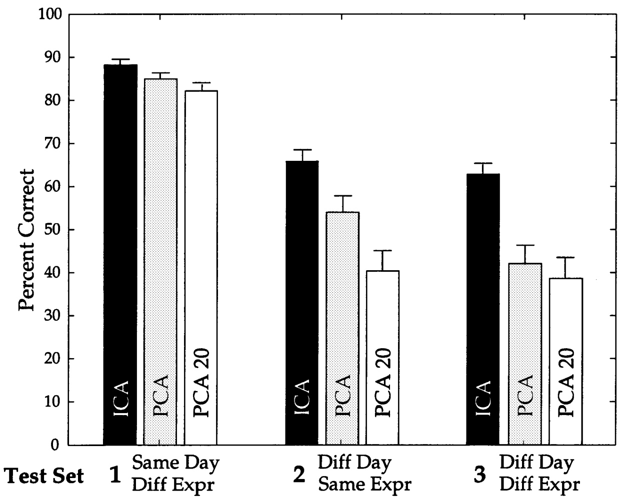


Fig. 10. Percent correct face recognition for the ICA representation, Architecture I, using 200 ICs, the PCA representation using 200 PCs, and the PCA representation using 20 PCs. Groups are performances for Test Set 1, Test Set 2, and Test Set 3. Error bars are one standard deviation of the estimate of the success rate for a Bernoulli distribution.

over all three test sets ($Z = 2.5$, $p < 0.05$) for Test Sets 1 and 2, and ($Z = 2.4 < 0.05$), $p < 0.05$) for Test Set 3.

Recognition performance using different numbers of ICs was also examined by performing ICA on 20 to 200 image mixtures in steps of 20. Best performance was obtained by separating 200 ICs. In general, the more ICs were separated, the better the recognition performance. The basis images also became increasingly spatially local as the number of separated components increased.

B. Subspace Selection

When all 200 components were retained, then PCA and ICA were working in the same subspace. However, as illustrated in Fig. 2, when subsets of axes are selected, then ICA chooses a different subspace from PCA. The full benefit of ICA may not be tapped until ICA-defined subspaces are explored.

Face recognition performances for the PCA and ICA representations were next compared by selecting subsets of the 200 components by class discriminability. Let $\overline{x}$ be the overall mean of a coefficient b_k across all faces, and $\overline{x}_j$ be the mean for person j. For both the PCA and ICA representations, we calculated the ratio of between-class to within-class variability r for each coefficient

$$r = \frac{\sigma_{\text{between}}}{\sigma_{\text{within}}} \tag{14}$$

where $\sigma_{\text{between}} = \sum_j (\overline{x}_j - \overline{x})^2$ is the variance of the j class means, and $\sigma_{\text{within}} = \sum_j \sum_i (x_{ij} - \overline{x}_j)^2$ is the sum of the variances within each class.

The class discriminability analysis was carried out using the 43 subjects for which four frontal view images were available. The ratios r were calculated separately for each test set, excluding the test images from the analysis. Both the PCA and ICA coefficients were then ordered by the magnitude of r. (Fig. 11 top) compares the discriminability of the ICA coefficients to the PCA coefficients. The ICA coefficients consistently had greater class discriminability than the PCA coefficients.

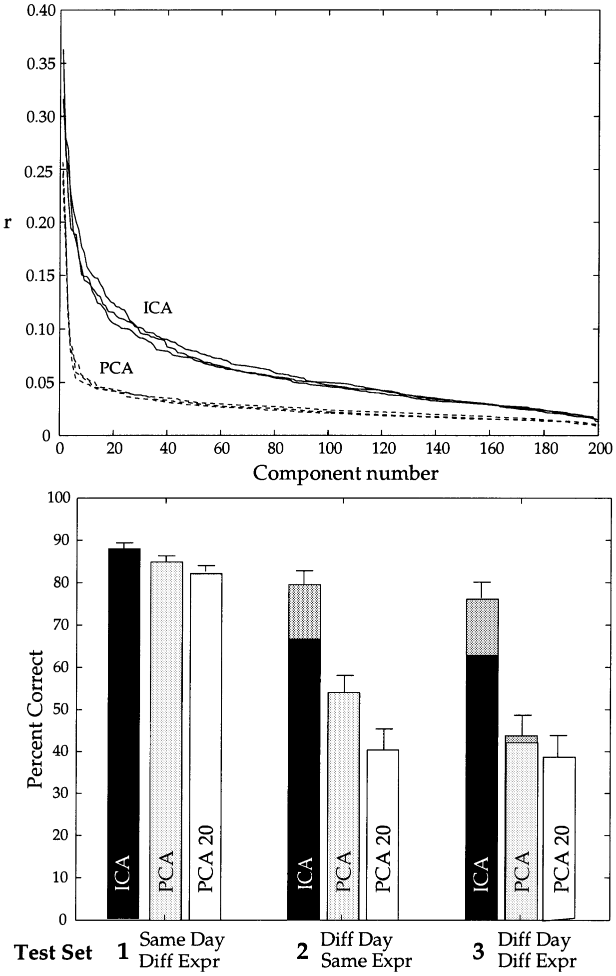


Fig. 11. Selection of components by class discriminability, Architecture II. Top: Discriminability of the ICA coefficients (solid lines) and discriminability of the PCA components (dotted lines) for the three test cases. Components were sorted by the magnitude of r. Bottom: Improvement in face recognition performance for the ICA and PCA representations using subsets of components selected by the class discriminability r. The improvement is indicated by the gray segments at the top of the bars.

Face classification performance was compared using the k most discriminable components of each representation. (Fig. 11 bottom) shows the best classification performance obtained for the PCA and ICA representations, which was with the 60 most discriminable components for the ICA representation, and the 140 most discriminable components for the PCA representation. Selecting subsets of coefficients by class discriminability improved the performance of the ICA representation, but had little effect on the performance of the PCA representation. The ICA representation again outperformed the PCA representation. The difference in recognition performance between the ICA and PCA representations was significant for Test Set 2 and Test Set 3, the two conditions that required recognition of images collected on a different day from the training set ($Z = 2.9, p < 0.05$; $Z = 3.4, p < 0.01$), respectively, when both subspaces were selected under the criterion of class discriminability. Here, the ICA-defined subspace encoded more information about facial identity than PCA-defined subspace.

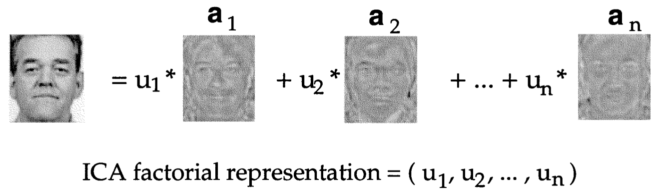


Fig. 12. The factorial code representation consisted of the independent coefficients, $\mathbf{u}$, for the linear combination of basis images in A that comprised each face image $\mathbf{x}$.

V. Architecture II: A Factorial Face Code

The goal in Architecture I was to use ICA to find a set of spatially independent basis images. Although the basis images obtained in that architecture are approximately independent, the coefficients that code each face are not necessarily independent. Architecture II uses ICA to find a representation in which the coefficients used to code images are statistically independent, i.e., a factorial face code. Barlow and Atick have discussed advantages of factorial codes for encoding complex objects that are characterized by high-order combinations of features [2], [5]. These include fact that the probability of any combination of features can be obtained from their marginal probabilities.

To achieve this goal, we organize the data matrix X so that rows represent different pixels and columns represent different images. [See (Fig. 3 right)]. This corresponds to treating the columns of $A \triangleq W_I^{-1}$ as a set of basis images. (See Fig. 5). The ICA representations are in columns of $U = W_I X$. Each column of U contains the coefficients of the basis images in A for reconstructing each image in X (Fig. 12). ICA attempts to make the outputs, U, as independent as possible. Hence, U is a factorial code for the face images. The representational code for test images is obtained by

$$W_I X_{\text{test}} = U_{\text{test}} \tag{15}$$

where X_{test} is the zero-mean[6] matrix of test images, and W_I is the weight matrix found by performing ICA on the training images.

In order to reduce the dimensionality of the input, instead of performing ICA directly on the 3000 image pixels, ICA was performed on the first 200 PCA coefficients of the face images. The first 200 PCs accounted for over 98% of the variance in the images. These coefficients R_{200}^T comprised the columns of the input data matrix, where each coefficient had zero mean. The Architecture II representation for the training images was therefore contained in the columns of U, where

$$W_I R_{200}^T = U. \tag{16}$$

The ICA weight matrix W_I was 200×200, resulting in 200 coefficients in U for each face image, consisting of the outputs of each of the ICA filters.[7] The architecture II representation for test images was obtained in the columns of U_{test} as follows:

$$W_I {R_{\text{test}}}^T = U_{\text{test}}. \tag{17}$$

The basis images for this representation consisted of the columns of $A \triangleq W_I^{-1}$. A sample of the basis images is shown

[6]Here, each pixel has zero mean.

[7]An image filter $f(\mathbf{x})$ is defined as $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$.

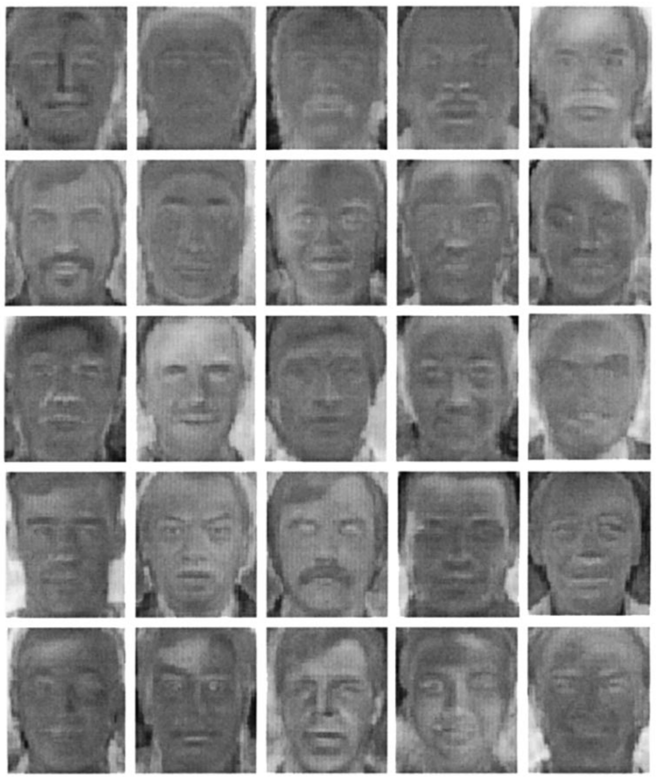


Fig. 13. Basis images for the ICA-factorial representation (columns of $A \triangleq W_I^{-1}$) obtained with Architecture II.

in Fig. 13, where the PC reconstruction P_{200}^A was used to visualize them. In this approach, each column of the mixing matrix W^{-1} found by ICA attempts to get close to a cluster of images that look similar across pixels. Thus, this approach tends to generate basis images that look more face-like than the basis images generated by PCA, in that the bases found by ICA will average only images that look alike. Unlike the ICA output U, the algorithm does not force the columns of A to be either sparse or independent. Indeed, the basis images in A have more global properties than the basis images in the ICA output of Architecture I shown in Fig. 8.

A. Face Recognition Performance: Architecture II

Face recognition performance was again evaluated by the nearest neighbor procedure using cosines as the similarity measure. Fig. 14 compares the face recognition performance using the ICA factorial code representation obtained with Architecture II to the independent basis representation obtained with Architecture I and to the PCA representation, each with 200 coefficients. Again, there was a trend for the ICA-factorial representation (ICA2) to outperform the PCA representation for recognizing faces across days. The difference in performance for Test Set 2 is significant ($Z = 2.7$, $p < 0.01$). There was no significant difference in the performances of the two ICA representations.

Class discriminability of the 200 ICA factorial coefficients was calculated according to (14). Unlike the coefficients in the independent basis representation, the ICA-factorial coefficients did not differ substantially from each other according to discriminability r. Selection of subsets of components for the

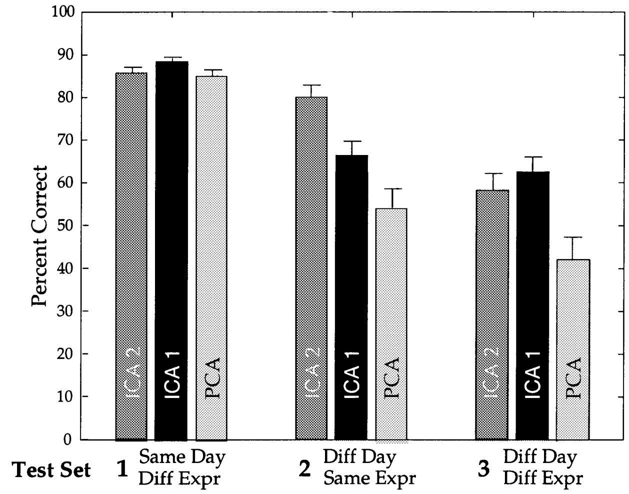


Fig. 14. Recognition performance of the factorial code ICA representation (ICA2) using all 200 coefficients, compared to the ICA independent basis representation (ICA1), and the PCA representation, also with 200 coefficients.

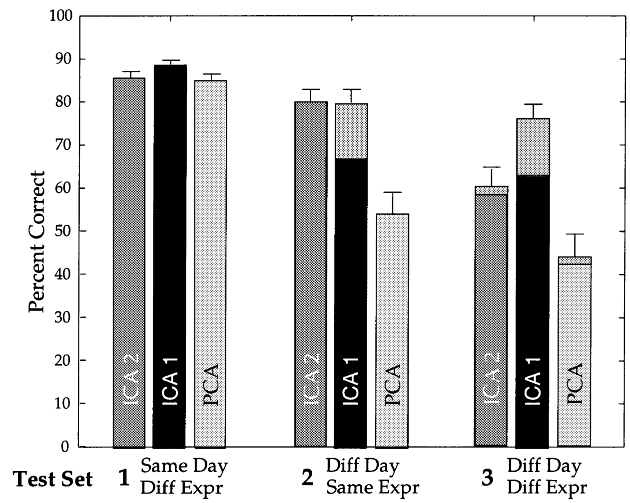


Fig. 15. Improvement in recognition performance of the two ICA representations and the PCA representation by selecting subsets of components by class discriminability. Gray extensions show improvement over recognition performance using all 200 coefficients.

representation by class discriminability had little effect on the recognition performance using the ICA-factorial representation (see Fig. 15). The difference in performance between ICA1 and ICA2 for Test Set 3 following the discriminability analysis just misses significance ($Z = 1.88$, $p = 0.06$).

In this implementation, we separated 200 components using 425 samples, which was a bare minimum. Test images were not used to learn the ICs, and thus our recognition results were not due to overlearning. Nevertheless, in order to determine whether the findings were an artifact due to small sample size, recognition performances were also tested after separating 85 rather than 200 components and, hence, estimating fewer weight parameters. The same overall pattern of results was obtained when 85 components were separated. Both ICA representations significantly outperformed the PCA representation on Test Sets 2 and 3. With 85 ICs, ICA1 obtained 87%, 62%, 58% correct performance, respectively, on Test Sets 1, 2, and 3, ICA2 obtained 85%, 76%, and 56% correct performance, whereas PCA

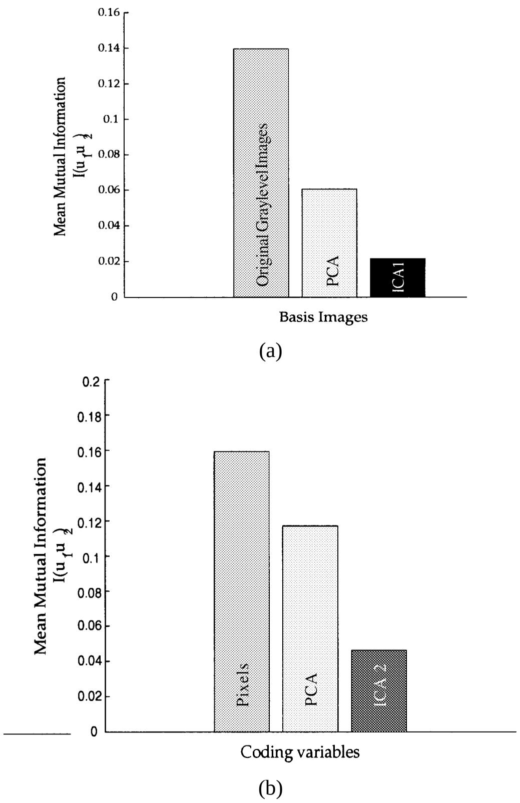


Fig. 16. Pairwise mutual information. (a) Mean mutual information between basis images. Mutual information was measured between pairs of gray-level images, PC images, and independent basis images obtained by Architecture I. (b) Mean mutual information between coding variables. Mutual information was measured between pairs of image pixels in gray-level images, PCA coefficients, and ICA coefficients obtained by Architecture II.

obtained 85%, 56%, and 44% correct, respectively. Again, as found for 200 separated components, selection of subsets of components by class discriminability improved the performance of ICA1 to 86%, 78%, and 65%, respectively, and had little effect on the performances with the PCA and ICA2 representations. This suggests that the results were not simply an artifact due to small sample size.

VI. Examination of the ICA Representations

A. Mutual Information

A measure of the statistical dependencies of the face representations was obtained by calculating the mean mutual information between pairs of 50 basis images. Mutual information was calculated as

$$I(\mathbf{U}_1, \mathbf{U}_2) = H(\mathbf{U}_1) + H(\mathbf{U}_2) - H(\mathbf{U}_1, \mathbf{U}_2) \qquad (18)$$

where $H(\mathbf{U}_i) = -E\left(\log(P_{U_i}(\mathbf{U}_i))\right)$.

Fig. 16 (a) compares the mutual information between *basis images* for the original gray-level images, the PC basis images, and the ICA basis images obtained in Architecture I. Principal component (PC) images are uncorrelated, but there are remaining high-order dependencies. The information maximization algorithm decreased these residual dependencies by more than 50%. The remaining dependence may be due to a mismatch between the logistic transfer function employed in the learning rule and the cumulative density function of the independent sources, the presence of sub-Gaussian sources, or the large number of free parameters to be estimated relative to the number of training images.

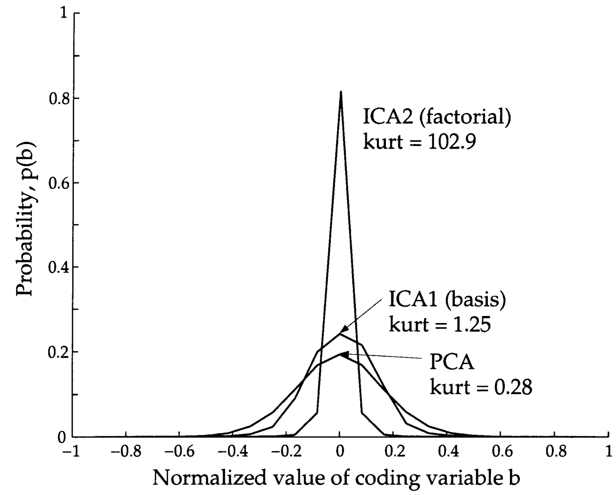


Fig. 17. Kurtosis (sparseness) of ICA and PCA representations.

Fig. 16 (b) compares the mutual information between the *coding variables* in the ICA factorial representation obtained with Architecture II, the PCA representation, and gray-level images. For gray-level images, mutual information was calculated between pairs of pixel locations. For the PCA representation, mutual information was calculated between pairs of PC coefficients, and for the ICA factorial representation, mutual information was calculated between pairs of coefficients u. Again, there were considerable high-order dependencies remaining in the PCA representation that were reduced by more than 50% by the information maximization algorithm. The ICA representations obtained in these simulations are most accurately described not as "independent," but as "redundancy reduced," where the redundancy is less than half that in the PC representation.

B. Sparseness

Field [19] has argued that sparse distributed representations are advantageous for coding visual stimuli. Sparse representations are characterized by highly kurtotic response distributions, in which a large concentration of values are near zero, with rare occurrences of large positive or negative values in the tails. In such a code, the redundancy of the input is transformed into the redundancy of the response patterns of the the individual outputs. Maximizing sparseness without loss of information is equivalent to the minimum entropy codes discussed by Barlow [5].[8]

Given the relationship between sparse codes and minimum entropy, the advantages for sparse codes as outlined by Field [19] mirror the arguments for independence presented by Barlow [5]. Codes that minimize the number of active neurons can be useful in the detection of suspicious coincidences. Because a nonzero response of each unit is relatively rare, high-order relations become increasingly rare, and therefore, more informative when they are present in the stimulus. Field

[8]Information maximization is consistent with minimum entropy coding. By maximizing the *joint* entropy of the output, the entropies of the *individual* outputs tend to be minimized.

Fig. 18. Recognition successes and failures. {left) Two face image pairs which both ICA algorithms correctly recognized. (right) Two face image pairs that were misidentified by both ICA algorithms. Images from the FERET face database were reprinted with permission from J. Phillips.

contrasts this with a compact code such as PCs, in which a few units have a relatively high probability of response, and therefore, high-order combinations among this group are relatively common. In a sparse distributed code, different objects are represented by which units are active, rather than by how much they are active. These representations have an added advantage in signal-to-noise, since one need only determine which units are active without regard to the precise level of activity. An additional advantage of sparse coding for face representations is storage in associative memory systems. Networks with sparse inputs can store more memories and provide more effective retrieval with partial information [10], [47].

The probability densities for the values of the coefficients of the two ICA representations and the PCA representation are shown in Fig. 17. The sparseness of the face representations were examined by measuring the kurtosis of the distributions. Kurtosis is defined as the ratio of the fourth moment of the distribution to the square of the second moment, normalized to zero for the Gaussian distribution by subtracting 3

$$\text{kurtosis} = \frac{\sum_i (b_i - \bar{b})^4}{\left(\sum_i (b_i - \bar{b})^2\right)^2} - 3. \tag{19}$$

The kurtosis of the PCA representation was measured for the PC coefficients. The PCs of the face images had a kurtosis of 0.28. The coefficients, b, of the independent basis representation from Architecture I had a kurtosis of 1.25. Although the basis images in Architecture I had a sparse distribution of gray-level values, the face coefficients with respect to this basis were not sparse. In contrast, the coefficients u of the ICA factorial code representation from Architecture II were highly kurtotic at 102.9.

VII. Combined ICA Recognition System

Given that the two ICA representations gave similar recognition performances, we examined whether the two representations gave similar patterns of errors on the face images. There was a significant tendency for the two algorithms to misclassify the same images. The probability that the ICA-factorial representation (ICA2) made an error given that the ICA1 representation made an error was 0.72, 0.88, and 0.89, respectively, for the three test sets. These conditional error rates were significantly higher than the marginal error rates ($Z = 7.4, p < 0.001$; $Z = 3.4, p < 0.001$; $Z = 2.8, p < 0.01$), respectively. Examples of successes and failures of the two algorithms are shown in Fig. 18.

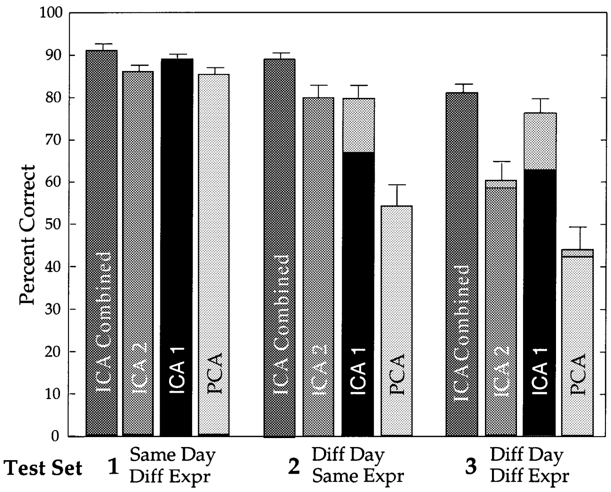


Fig. 19. Face recognition performance of the ocmbined ICA classifier, compared to the individual classifiers for ICA1, ICA2, and PCA.

When the two algorithms made errors, however, they did not assign the same incorrect identity. Out of a total of 62 common errors between the two systems, only once did both algorithms assign the same incorrect identity. The two representations can, therefore, used in conjunction to provide a reliability measure, where classifications are accepted only if both algorithms gave the same answer. The ICA recognition system using this reliability criterion gave a performance of 100%, 100%, and 97% for the three test sets, respectively, which is an overall classification performance of 99.8%. 400 out of the total of 500 test images met criterion.

Because the confusions made by the two algorithms differed, a combined classifier was employed in which the similarity between a test image and a gallery image was defined as $c_1 + c_2$, where c_1 and c_2 correspond to the similarity measure c in (12) for ICA1 and ICA2, respectively. Class discriminability analysis was carried out on ICA1 and ICA2 before calculating c_1 and c_2. Performance of the combined classifier is shown in Fig. 19. The combined classifier improved performance to 91.0%, 88.9%, and 81.0% for the three test cases, respectively. The difference in performance between the combined ICA classifier and PCA was significant for all three test sets ($Z = 2.7, p < 0.01$; $Z = 3.7$, $p < 0.001$; $Z = 3.7$; $p < 0.001$).

VIII. Discussion

Much of the information that perceptually distinguishes faces is contained in the higher order statistics of the images, i.e., the phase spectrum. The basis images developed by PCA depend only on second-order images statistics and, thus, it is desirable to find generalizations of PCA that are sensitive to higher order image statistics. In this paper, we explored one such generalization: Bell and Sejnowski's ICA algorithm. We explored two different architectures for developing image representations of faces using ICA. Architecture I treated images as random variables and pixels as random trials. This architecture was related to the one used by Bell and Sejnowski to separate mixtures of auditory signals into independent sound sources. Under this architecture, ICA found a basis set of statistically independent images. The images in this basis set were sparse and localized in space, resembling facial features. Architecture II treated pixels as random variables and images as random trials. Under this architecture, the image coefficients were approximately independent, resulting in a factorial face code.

Both ICA representations outperformed the "eigenface" representation [57], which was based on PC analysis, for recognizing images of faces sampled on a different day from the training images. A classifier that combined the two ICA representations outperformed eigenfaces on all test sets. Since ICA allows the basis images to be nonorthogonal, the angles and distances between images differ between ICA and PCA. Moreover, when subsets of axes are selected, ICA defines a different subspace than PCA. We found that when selecting axes according to the criterion of class discriminability, ICA-defined subspaces encoded more information about facial identity than PCA-defined subspaces.

ICA representations are designed to maximize information transmission in the presence of noise and, thus, they may be more robust to variations such as lighting conditions, changes in hair, make-up, and facial expression, which can be considered forms of noise with respect to the main source of information in our face database: the person's identity. The robust recognition across different days is particularly encouraging, since most applications of automated face recognition contain the noise inherent to identifying images collected on a different day from the sample images.

The purpose of the comparison in this paper was to examine ICA and PCA-based representations under identical conditions. A number of methods have been presented for enhancing recognition performance with eigenfaces (e.g., [41] and [51]). ICA representations can be used in place of eigenfaces in these techniques. It is an open question as to whether these techniques would enhance performance with PCA and ICA equally, or whether there would be interactions between the type of enhancement and the representation.

A number of research groups have independently tested the ICA representations presented here and in [9]. Liu and Wechsler [35], and Yuen and Lai [61] both supported our findings that ICA outperformed PCA. Moghaddam [41] employed Euclidean distance as the similarity measure instead of cosines. Consistent with our findings, there was no significant difference between PCA and ICA using Euclidean distance as the similarity measure. Cosines were not tested in that paper. A thorough comparison of ICA and PCA using a large set of similarity measures was recently conducted in [17], and supported the advantage of ICA for face recognition.

In Section V, ICA provided a set of statistically independent coefficients for coding the images. It has been argued that such a factorial code is advantageous for encoding complex objects that are characterized by high-order combinations of features, since the prior probability of any combination of features can be obtained from their individual probabilities [2], [5]. According to the arguments of both Field [19] and Barlow [5], the ICA-factorial representation (Architecture II) is a more optimal object representation than the Architecture I representation given its sparse, factorial properties. Due to the difference in architecture, the ICA-factorial representation always had fewer training samples to estimate the same number of free parameters as the Architecture I representation. Fig. 16 shows that the residual dependencies in the ICA-factorial representation were higher than in the Architecture I representation. The ICA-factorial representation may prove to have a greater advantage given a much larger training set of images. Indeed, this prediction has born out in recent experiments with a larger set of FERET face images [17]. It also is possible that the factorial code representation may prove advantageous with more powerful recognition engines than nearest neighbor on cosines, such as a Bayesian classifier. An image set containing many more frontal view images of each subject collected on different days will be needed to test that hypothesis.

In this paper, the number of sources was controlled by reducing the dimensionality of the data through PCA prior to performing ICA. There are two limitations to this approach [55]. The first is the reverse dimensionality problem. It may not be possible to linearly separate the independent sources in smaller subspaces. Since we retained 200 dimensions, this may not have been a serious limitation of this implementation. Second, it may not be desirable to throw away subspaces of the data with low power such as the higher PCs. Although low in power, these subspaces may contain ICs, and the property of the data we seek is independence, not amplitude. Techniques have been proposed for separating sources on projection planes without discarding any ICs of the data [55]. Techniques for estimating the number of ICs in a dataset have also recently been proposed [26], [40].

The information maximization algorithm employed to perform ICA in this paper assumed that the underlying "causes" of the pixel gray-levels in face images had a super-Gaussian (peaky) response distribution. Many natural signals, such as sound sources, have been shown to have a super-Gaussian distribution [11]. We employed a logistic source model which has shown in practice to be sufficient to separate natural signals with super-Gaussian distributions [11]. The underlying "causes" of the pixel gray-levels in the face images are unknown, and it is possible that better results could have been obtained with other source models. In particular, any sub-Gaussian sources would have remained mixed. Methods for separating sub-Gaussian sources through information maximization have been developed [30]. A future direction of this research is to examine sub-Gaussian components of face images.

The information maximization algorithm employed in this work also assumed that the pixel values in face images were generated from a linear mixing process. This linear approximation has been shown to hold true for the effect of lighting on face images [21]. Other influences, such as changes in pose and expression may be linearly approximated only to a limited extent. Nonlinear ICA in the absence of prior constraints is an ill-conditioned problem, but some progress has been made by assuming a linear mixing process followed by parametric nonlinear functions [31], [59]. An algorithm for nonlinear ICA based on kernel methods has also recently been presented [4]. Kernel methods have already shown to improve face recognition performance with PCA and Fisherfaces [60]. Another future direction of this research is to examine nonlinear ICA representations of faces.

Unlike PCA, the ICA using Architecture I found a spatially local face representation. Local feature analysis (LFA) [50] also finds local basis images for faces, but using second-order statistics. The LFA basis images are found by performing whitening (4) on the PC axes, followed by a rotation to topographic correspondence with pixel location. The LFA kernels are not sensitive to the high-order dependencies in the face image ensemble, and in tests to date, recognition performance with LFA kernels has not significantly improved upon PCA [16]. Interestingly, downsampling methods based on sequential information maximization significantly improve performance with LFA [49].

ICA outputs using Architecture I were sparse in space (within image across pixels) while the ICA outputs using Architecture II were sparse across images. Hence Architecture I produced local basis images, but the face codes were not sparse, while Architecture II produced sparse face codes, but with holistic basis images. A representation that has recently appeared in the literature, nonnegative matrix factorization (NMF) [28], produced local basis images and sparse face codes.[9] While this representation is interesting from a theoretical perspective, it has not yet proven useful for recognition. Another innovative face representation employs products of experts in restricted Boltzmann machines (RBMs). This representation also finds local features when nonnegative weight constraints are employed [56]. In experiments to date, RBMs outperformed PCA for recognizing faces across changes in expression or addition/removal of glasses, but performed more poorly for recognizing faces across different days. It is an open question as to whether sparseness and local features are desirable objectives for face recognition in and of themselves. Here, these properties emerged from an objective of independence.

Capturing more likelihood may be a good principle for generating unsupervised representations which can be later used for classification. As mentioned in Section II, PCA and ICA can be derived as generative models of the data, where PCA uses Gaussian sources, and ICA typically uses sparse sources. It has been shown that for many natural signals, ICA is a better model in that it assigns higher likelihood to the data than PCA [32]. The ICA basis dimensions presented here may have captured more likelihood of the face images than PCA, which provides a possible explanation for the superior performance of ICA for face recognition in this study.

The ICA representations have a degree of biological relevance. The information maximization learning algorithm was developed from the principle of optimal information transfer in neurons with sigmoidal transfer functions. It contains a Hebbian correlational term between the nonlinearly transformed outputs and weighted feedback from the linear outputs [12]. The biological plausibility of the learning algorithm, however, is limited by fact that the learning rule is nonlocal. Local learning rules for ICA are presently under development [34], [38].

The principle of independence, if not the specific learning algorithm employed here [12], may have relevance to face and object representations in the brain. Barlow [5] and Atick [2] have argued for redundancy reduction as a general coding strategy in the brain. This notion is supported by the findings of Bell and Sejnowski [12] that image bases that produce independent outputs from natural scenes are local oriented spatially opponent filters similar to the response properties of V1 simple cells. Olshausen and Field [43], [44] obtained a similar result with a sparseness objective, where there is a close information theoretic relationship between sparseness and independence [5], [12]. Conversely, it has also been shown that Gabor filters, which closely model the responses of V1 simple cells, separate high-order dependencies [18], [19], [54]. (See [6] for a more detailed discussion). In support of the relationship between Gabor filters and ICA, the Gabor and ICA Architecture I representations significantly outperformed more than eight other image representations on a task of facial expression recognition, and performed equally well to each other [8], [16]. There is also psychophysical support for the relevance of independence to face representations in the brain. The ICA Architecture I representation gave better correspondence with human perception of facial similarity than both PCA and nonnegative matrix factorization [22].

Desirable filters may be those that are adapted to the patterns of interest and capture interesting structure [33]. The more the dependencies that are encoded, the more structure that is learned. Information theory provides a means for capturing interesting structure. Information maximization leads to an efficient code of the environment, resulting in more learned structure. Such mechanisms predict neural codes in both vision [12], [43], [58] and audition [32]. The research presented here found that face representations in which high-order dependencies are separated into individual coefficients gave superior recognition performance to representations which only separate second-order redundancies.

[9]Although the NMF codes were sparse, they were not a minimum entropy code (an independent code) as the objective function did not maximize sparseness while preserving information.

Acknowledgment

The authors are grateful to M. Lades, M. McKeown, M. Gray, and T.-W. Lee for helpful discussions on this topic, and valuable comments on earlier drafts of this paper.

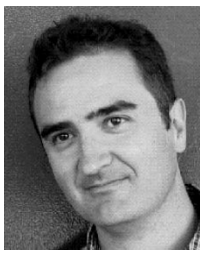
References

[1] S. Amari, A. Cichocki, and H. H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1996, vol. 8.
[2] J. J. Atick, “Could information theory provide an ecological theory of sensory processing?,” *Network*, vol. 3, pp. 213–251, 1992.
[3] J. J. Atick and A. N. Redlich, “What does the retina know about natural scenes?,” *Neural Comput.*, vol. 4, pp. 196–210, 1992.
[4] F. R. Bach and M. I. Jordan, “Kernel independent component analysis,” *J. Machine Learning Res.*, vol. 3, pp. 1–48, 2002.
[5] H. B. Barlow, “Unsupervised learning,” *Neural Comput.*, vol. 1, pp. 295–311, 1989.
[6] M. S. Bartlett, *Face Image Analysis by Unsupervised Learning*. Boston, MA: Kluwer, 2001, vol. 612, Kluwer International Series on Engineering and Computer Science.
[7] ——, “Face Image Analysis by Unsupervised Learning and Redundancy Reduction,” Ph.D. dissertation, Univ. California-San Diego, La Jolla, 1998.
[8] M. S. Bartlett, G. L. Donato, J. R. Movellan, J. C. Hager, P. Ekman, and T. J. Sejnowski, “Image representations for facial expression coding,” in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Muller, Eds. Cambridge, MA: MIT Press, 2000, vol. 12.

[9] M. S. Bartlett, H. M. Lades, and T. J. Sejnowski, "Independent component representations for face recognition," in *Proc. SPIE Symp. Electon. Imaging: Science Technology—Human Vision and Electronic Imaging III*, vol. 3299, T. Rogowitz and B. Pappas, Eds., San Jose, CA, 1998, pp. 528–539.
[10] E. B. Baum, J. Moody, and F. Wilczek, "Internal representaions for associative memory," *Biol. Cybern.*, vol. 59, pp. 217–228, 1988.
[11] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.
[12] ——, "The independent components of natural scenes are edge filters," *Vision Res.*, vol. 37, no. 23, pp. 3327–3338, 1997.
[13] A. Cichocki, R. Unbehauen, and E. Rummert, "Robust learning algorithm for blind separation of signals," *Electron. Lett.*, vol. 30, no. 7, pp. 1386–1387, 1994.
[14] P. Comon, "Independent component analysis—A new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.
[15] G. Cottrell and J. Metcalfe, "Face, gender and emotion recognition using holons," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed. San Mateo, CA: Morgan Kaufmann, 1991, vol. 3, pp. 564–571.
[16] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 974–989, Oct. 1999.
[17] B. A. Draper, K. Baek, M. S. Bartlett, and J. R. Beveridge, "Recognizing faces with PCA and ICA," Comput. Vision Image Understanding (Special Issue on Face Recognition), 2002, submitted for publication.
[18] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, pp. 2379–94, 1987.
[19] ——, "What is the goal of sensory coding?," *Neural Comput.*, vol. 6, pp. 559–601, 1994.
[20] M. Girolami, *Advances in Independent Component Analysis*. Berlin, Germany: Springer-Verlag, 2000.
[21] P. Hallinan, "A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1995.
[22] P. Hancock, "Alternative representations for faces," in *British Psych. Soc., Cognitive Section*. Essex, U.K.: Univ. Essex, 2000.
[23] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neurosci.*, vol. 9, pp. 181–197, 1992.
[24] G. Hinton and T. Shallice, "Lesioning an attractor network: Investigations of acquired dyslexia," *Psych. Rev.*, vol. 98, no. 1, pp. 74–95, 1991.
[25] C. Jutten and J. Herault, "Blind separation of sources i. an adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.
[26] H. Lappalainen and J. W. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*, M. Girolami, Ed. New York: Springer-Verlag, 2000, pp. 76–92.
[27] S. Laughlin, "A simple coding procedure enhances a neuron's information capacity," *Z. Naturforsch.*, vol. 36, pp. 910–912, 1981.
[28] D. D. Lee and S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[29] T.-W. Lee, *Independent Component Analysis: Theory and Applications*. Boston, MA: Kluwer, 1998.
[30] T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Comput.*, vol. 11, no. 2, pp. 417–41, 1999.
[31] T.-W. Lee, B. U. Koehler, and R. Orglmeister, "Blind source separation of nonlinear mixing models," in *Proc. IEEE Int. Workshop Neural Networks Signal Processing*, Sept. 1997, pp. 406–415.
[32] M. Lewicki and B. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Amer. A*, vol. 16, no. 7, pp. 1587–601, 1999.
[33] M. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–65, 2000.
[34] J. Lin, D. G. Grier, and J. Cowan, "Source separation and density estimation by faithful equivariant som," in *Advances in Neural Information Processing Systems*, M. Mozer, M. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 536–541.
[35] C. Liu and H. Wechsler, "Comparative assessment of independent component analysis (ICA) for face recognition," presented at the Int. Conf. Audio Video Based Biometric Person Authentication, 1999.
[36] D. J. C. MacKay, Maximum Likelihood and Covariant Algorithms for Independent Component Analysis: , 1996.
[37] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 145–151.
[38] T. K. Marks and J. R. Movellan, "Diffusion networks, products of experts, and factor analysis," in *Proc. 3rd Int. Conf. Independent Component Anal. Signal Separation*, 2001.
[39] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fMRI by decomposition into independent spatial components," *Human Brain Mapping*, vol. 6, no. 3, pp. 160–88, 1998.
[40] J. W. Miskin and D. J. C. MacKay, *Ensemble Learning for Blind Source Separation ICA: Principles and Practice*. Cambridge, U.K.: Cambridge Univ. Press, 2001.
[41] B. Moghaddam, "Principal manifolds and Bayesian subspaces for visual recognition," presented at the Int. Conf. Comput. Vision, 1999.
[42] J.-P. Nadal and N. Parga, "Non-linear neurons in the low noise limit: A factorial code maximizes information transfer," *Network*, vol. 5, pp. 565–581, 1994.
[43] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.
[44] ——, "Natural image statistics and efficient coding," *Network: Comput. Neural Syst.*, vol. 7, no. 2, pp. 333–340, 1996.
[45] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, pp. 529–541, 1981.
[46] A. O'Toole, K. Deffenbacher, D. Valentin, and H. Abdi, "Structural aspects of face recognition and the other race effect," *Memory Cognition*, vol. 22, no. 2, pp. 208–224, 1994.
[47] G. Palm, "On associative memory," *Biol. Cybern.*, vol. 36, pp. 19–31, 1980.
[48] B. A. Pearlmutter and L. C. Parra, "A context-sensitive generalization of ICA," in *Advances in Neural Information Processing Systems*, Mozer, Jordan, and Petsche, Eds. Cambridge, MA: MIT Press, 1996, vol. 9.
[49] P. S. Penev, "Redundancy and dimensionality reduction in sparse-distributed representations of natural objects in terms of their local features," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001.
[50] P. S. Penev and J. J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Comput. Neural Syst.*, vol. 7, no. 3, pp. 477–500, 1996.
[51] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 1994, pp. 84–91.
[52] P. J. Phillips, H. Wechsler, J. Juang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image Vision Comput. J.*, vol. 16, no. 5, pp. 295–306, 1998.
[53] L. N. Piotrowski and F. W. Campbell, "A demonstration of the visual importance and flexibility of spatial-frequency, amplitude, and phase," *Perception*, vol. 11, pp. 337–346, 1982.
[54] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," presented at the 31st Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Nov. 2–5, 1997.
[55] J. V. Stone and J. Porrill, "Undercomplete Independent Component Analysis for Signal Separation and Dimension Reduction, Tech. Rep.," Dept. Psych., Univ. Sheffield, Sheffield, U.K., 1998.
[56] Y. W. Teh and G. E. Hinton, "Rate-coded restricted boltzmann machines for face recognition," in *Advances in Neural Information Processing Systems 13*, T. Leen, T. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001.
[57] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
[58] T. Wachtler, T.-W. Lee, and T. J. Sejnowski, "The chromatic structure of natural scenes," *J. Opt. Soc. Amer. A*, vol. 18, no. 1, pp. 65–77, 2001.
[59] H. H. Yang, S.-I. Amari, and A. Cichocki, "Nformation-theoretic approach to blind separation of sources in nonlinear mixture," *Signal Processing*, vol. 64, no. 3, pp. 291–3000, 1998.
[60] M. Yang, "Face recognition using kernel methods," in *Advances in Neural Information Processing Systems*, T. Diederich, S. Becker, and Z. Ghahramani, Eds., 2002, vol. 14.
[61] P. C. Yuen and J. H. Lai, "Independent component analysis of face images," presented at the IEEE Workshop Biologically Motivated Computer Vision, Seoul, Korea, 2000.

Marian Stewart Bartlett (M'99) received the B.S. degree in mathematics and computer science from Middlebury College, Middlebury, VT, in 1988 and the Ph.D. degree in cognitive science and psychology from the University of California-San Diego, La Jolla, in 1998. Her dissertation work was conducted with T. Sejnowski at the Salk Institute.

She is an Assistant Research Professor at the Institute for Neural Computation, University of California-San Diego. Her interests include approaches to image analysis through unsupervised learning, with a focus on face recognition and expression analysis. She is presently exploring probabilistic dynamical models and their application to facial expression analysis at the University of California-San Diego. She has also studied perceptual and cognitive processes with V.S. Ramachandran at the University of California-San Diego, the Cognitive Neuroscience Section of the National Institutes of Health, the Department of Brain and Cognitive Sciences at Massachusetts Institute of Technology, Cambridge, and the Brain and Perception Laboratory at the University of Bristol, U.K.



Javier R. Movellan (M'99) was born in Palencia, Spain, and received the B.S. degree from the Universidad Autonoma de Madrid, Madrid, Spain. He was a Fulbright Scholar at the University of California-Berkeley, Berkeley, and received the Ph.D. degree from the same university in 1989.

He was a Research Associate with Carnegie-Mellon University, Pittsburgh, PA, from 1989 to 1993, and an Assistant Professor with the Department of Cognitive Science, University of California-San Diego (USCD), La Jolla, from 1993 to 2001. He currently is a Research Associate with the Institute for Neural Computation and head of the Machine Perception Laboratory at UCSD. His research interests include the development of perceptual computer interfaces (i.e., system that recognize and react to natural speech commands, expressions, gestures, and body motions), analyzing the statistical structure of natural signals in order to help understand how the brain works, and the application of stochastic processes and probability theory to the study of the brain, behavior, and computation.

Terrence J. Sejnowski (S'83–SM'91–F'00) received the B.S. degree in physics from the Case Western Reserve University, Cleveland, OH, and the Ph.D. degree in physics from Princeton University, Princeton, NJ, in 1978.

In 1982, he joined the faculty of the Department of Biophysics at Johns Hopkins University, Baltimore, MD. He is an Investigator with the Howard Hughes Medical Institute and a Professor at The Salk Institute for Biological Studies, La Jolla, CA, where he directs the Computational Neurobiology Laboratory, and Professor of Biology at the University of California-San Diego, La Jolla. The long-range goal his research is to build linking principles from brain to behavior using computational models. This goal is being pursued with a combination of theoretical and experimental approaches at several levels of investigation ranging from the biophysical level to the systems level. The issues addressed by this research include how sensory information is represented in the visual cortex.

Dr. Sejnowski received the IEEE Neural Networks Pioneer Award in 2002.