# Face Modeling by Information Maximization[1]

Marian Stewart Bartlett
UC San Diego
marni@salk.edu

Javier R. Movellan
UC San Diego
movellan@mplab.ucsd.edu

Terrence J. Sejnowski
UC San Diego; Salk
Howard Hughes Medical Inst.
terry@salk.edu

**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 pair-wise 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 maximum information transfer through sigmoidal neurons [12]. 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 principal component analysis for recognizing faces across days and changes in expression. A computational neuroscience perspective on face modeling and information maximization is discussed.*

## 0.1 Introduction

Redundancy in the sensory input contains structural information about the environment. Horace 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. Barlow argues 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 [70], holons [16], and 'local feature analysis' [59] 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

[1] Based on "Face recognition by independent component analysis", by M.S. Bartlett, J.R. Movellan, & T.J. Sejnowski, which appeared in IEEE Transactions on Neural Networks 13(6) p. 1450-64, ©2002 IEEE.

analysis (ICA) [15] is one such generalization. A number of algorithms for performing ICA have been proposed. See [36, 23] for reviews. Here we employ an algorithm developed by Bell and Sejnowski [12, 13] 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 EEG signals [45], fMRI signals [46].

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 pixels as random variables and the images as outcomes[2] Matlab code for the ICA representations is available at http://mplab.ucsd.edu/˜marni.

Face recognition performance was tested using the FERET database [61]. Face recognition performances using the ICA representations were benchmarked by comparing them to performances using principal component analysis, which is equivalent to the "eigenfaces" representation [70, 60]. The two ICA representations were then combined in a single classifier.

## 0.2 Independent component analysis

There are a number of algorithms for performing ICA [29, 15, 14, 12]. We chose the infomax algorithm proposed by Bell and Sejnowski [12], which was derived from the principle of optimal information transfer in neurons with sigmoidal transfer functions [34]. The algorithm is motivated as follows: Let $\mathbf{X}$ be an n-dimensional random vector representing a distribution of inputs in the environment. (Here bold-face 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-dimensional random variable representing the outputs of n-neurons. Here each component of $f = (f_1, \cdots, f_n)$ is an invertible squashing function, mapping real numbers into the $[0, 1]$ interval. Typically the logistic function is used

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

The $\mathbf{U}_1, \cdots, \mathbf{U}_n$ variables are linear combinations of inputs and can be interpreted as presynaptic activations of n-neurons. The $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ variables can be interpreted as post-synaptic 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

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

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 [13]

$$\Delta W \propto \nabla_W H(\mathbf{Y}) W^T W = W + E(\mathbf{Y}'\mathbf{X}^T) W^T W \tag{3}$$

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 towards statistical independence. When the form of the nonlinear transfer function $f$ is the same as the cumulative density functions of the underlying independent components (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

[2] Preliminary versions of this work appear in [9, 7]. A longer discussion of unsupervised learning for face recognition appears in the following book [6].

in **U** [49, 13]. In practice, the logistic transfer function has been found sufficient to separate mixtures of natural signals with sparse distributions including sound sources [12].

The algorithm is speeded up by including a "sphering" step prior to learning [13]. The row means of **X** are subtracted, and then **X** is passed through the whitening matrix, $W_z$, which is twice the inverse principal square root[3] of the covariance matrix:

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

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,

$$W_I = WW_z. \tag{5}$$

MacKay [44] and Pearlmutter [57] showed that the ICA algorithm converges to the maximum likelihood estimate of $W^{-1}$ for the following generative model of the data

$$\mathbf{X} = W^{-1}\mathbf{S} \tag{6}$$

where $\mathbf{S} = (\mathbf{S}_1, \cdots, \mathbf{S}_n)'$ is a vector of independent random variables, called the sources, with cumulative distributions equal to $f_i$, i.e., 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 estimates of the sources that generated the data.

### 0.2.1 ICA and other statistical techniques

**ICA and PCA:** Principal component analysis (PCA) can be derived as a special case of ICA which uses Gaussian source models. In such case the mixing matrix $W$ is unidentifiable in the sense that there is an infinite number of equally good maximum likelihood solutions. Amongst all possible maximum likelihood solutions, PCA chooses an orthogonal matrix which is optimal in the following sense: (1) Regardless of the distribution of **X**, $\mathbf{U}_1$ is the linear combination of input that allows optimal linear reconstruction of the input in the mean square sense; (2) For $\mathbf{U}_1, \cdots \mathbf{U}_k$ fixed, $\mathbf{U}_{k+1}$ allows optimal linear reconstruction among the class of linear combinations of **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. In shift invariant databases (e.g. databases of natural images) the second-order statistics capture the amplitude spectrum of images but not their phase spectrum. The high order statistics capture the phase spectrum [22, 13]. For a given sample 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, 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 [54, 62]. 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 tail distributions [22, 12]. 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

[3] the unique square root for which every eigenvalue has nonnegative real part.

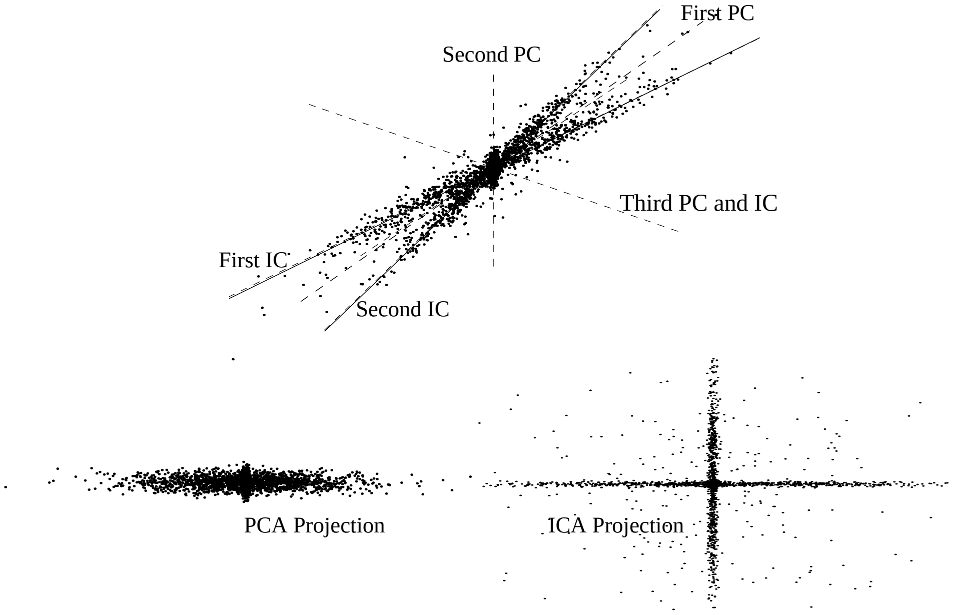

Figure 1: Top: Example 3-D data distribution and corresponding principal component and independent component axis. Each axis is a column of the mixing matrix $W^{-1}$ found by PCA or ICA. Note the PC axes are orthogonal while the IC axes are not. If only 2 components are allowed, ICA chooses a different 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 axis are non-orthogonal, relative distances between points are different in PCA than in ICA, as are the angles between points.

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.

Figure 1 illustrates these points with an example. The figure shows samples from a 3-dimensional 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 non-orthogonal, they change the relative distance between data points. This can be illustrated with the following example. Consider the three points: $x_1 = (4,0), x_2 = (0,10), x_3 = (10,10)$, and the following nonorthogonal basis set: $A = [1,1;0,1]$. The coordinates $y$ under the new basis set are defined by $x = Ay$ and thus $y = A^{-1}x$ where $A^{-1} = [1,-1;0,1]$. Thus the coordinates under the new basis set are $y_1 = (4,0), y_2 = (-10,10), y_3 = (0,10)$. Note in the standard coordinate system $x_1$ is closer to $x_2$ than to $x_3$. However in the new coordinate system $y_1$ is closer to $y_3$ than to $y_2$. In the old coordinate system the angle between $x_1$ and $x_3$ is the same as the angle between $x_2$ and $x_3$. However in the new coordinate system the angle between $y_1$ and $y_3$ is larger than the angle between $y_2$ and $y_3$.

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 illustrated in Figure 1 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 Figure 1, when only two dimensions are selected, PCA and ICA choose different sub-spaces.

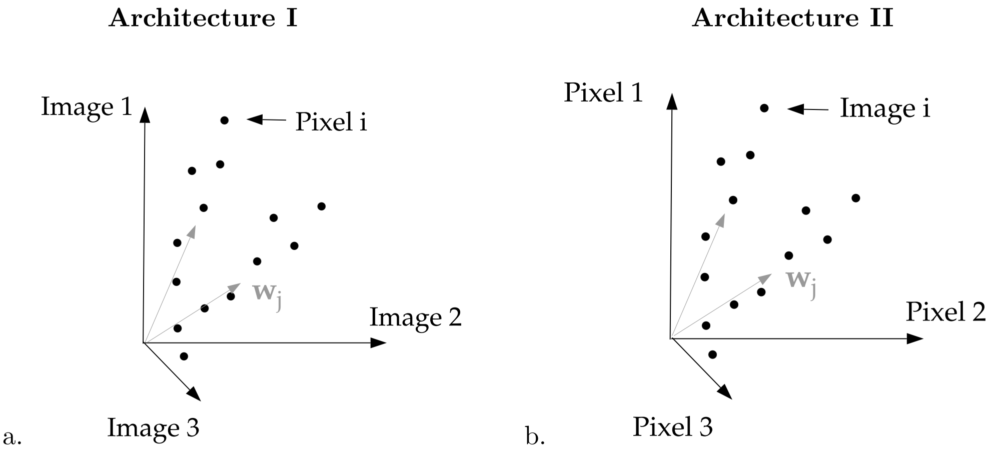

Figure 2: Two architectures for performing ICA on images. (a) Each pixel is plotted according to the grayvalue it takes on over a set of face images. ICA in architecture I finds weight vectors in the directions of statistical dependencies among the pixel locations. This defined a set of independent basis images. (b) Here, each image is an observation in a high dimensional space where the dimensions are the pixels. ICA in architecture II finds weight vectors in the directions of statistical dependencies among the face images. This defined a factorial face code.

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 [49]. 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 Figure 1 we found that if only 12 bits are allowed to represent the PCA and ICA coefficients, linear reconstructions based on ICA are 3dB 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 4 bits are allowed to represent the first 2 PCA and ICA coefficients, ICA reconstructions are 3dB 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.

When the sources models are sparse, ICA is closely related to the so called non-orthogonal "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 non-orthogonal "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. Figure 1 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 non-orthogonal PCA and to cluster analysis, however when the source models are sub-Gaussian, the relationship between these techniques is less clear. See [37] for a discussion of ICA in the context of sub-Gaussian sources.

### 0.2.2 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 $i^{th}$ 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 $\mathbf{X}_1, \cdots \mathbf{X}_{n_r}$ in which each column of $X$ is given probability mass $1/n_c$. Independence is then defined with respect to such 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.,

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

where $P$ is the empirical distribution as defined above.

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 Figure 2a. 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 & Sejnowski for sound source separation [12], for EEG analysis [45], and for fMRI [46]. The image synthesis model associated with this approach is illustrated in the top row of Figure 3.

2. We can transpose $X$ and organize our data so that images are in the columns of $X$. This approach is illustrated in Figure 2b. 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 & Sejnowski's work on the independent components of natural images [13]. The image synthesis model associated with this approach is illustrated in the bottom row of Figure 3.

## 0.3 Image data

The face images employed for this research were a subset of the FERET face database [61]. 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. The algorithms were trained on a single frontal view of each 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 1).

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 × 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.

## 0.4 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.,

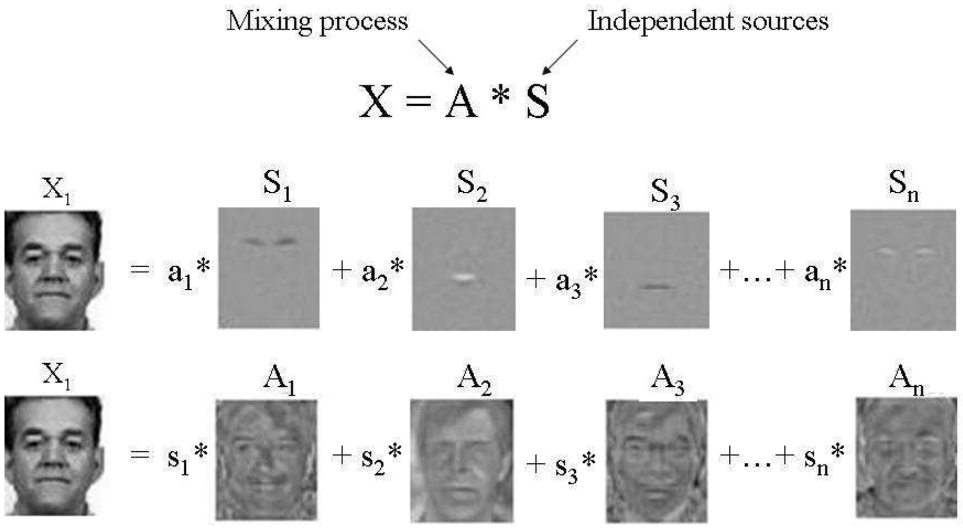
Mixing process
Independent sources
X = A * S
$X_1$
$S_1$
$S_2$
$S_3$
$S_n$
$= a_1* \quad + a_2* \quad + a_3* \quad +\ldots+ a_n*$
$X_1$
$A_1$
$A_2$
$A_3$
$A_n$
$= s_1* \quad + s_2* \quad + s_3* \quad +\ldots+ s_n*$

Figure 3: Image synthesis models for the two architectures. The ICA model decomposes images as **X**=**AS**, where **A** is a mixing matrix and **S** is a matrix of independent sources. In Architecture I (top), **S** contains the basis images and **A** contains the coefficients, whereas in Architecture II (bottom) **S** contains the coefficients and **A** contains the basis images for constructing each face image in **X**.

| **Image Set** | **Condition** | | **No. Images** |
|---|---|---|---|
| Training Set | Session I | 50% neutral 50% expr | 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 |

Table 1: Image sets used for training and testing.

$X$ has 425 rows and 3000 columns, and each image has zero mean.

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 Figure 4. These coordinates are contained in the mixing matrix $A \triangleq W_I^{-1}$.

The number of independent components 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 independent components. 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 independent components 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 independent component analysis of fMRI data [46], we chose for these linear combinations the first $m$ principal component eigenvectors of the image set. Principal component analysis 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)

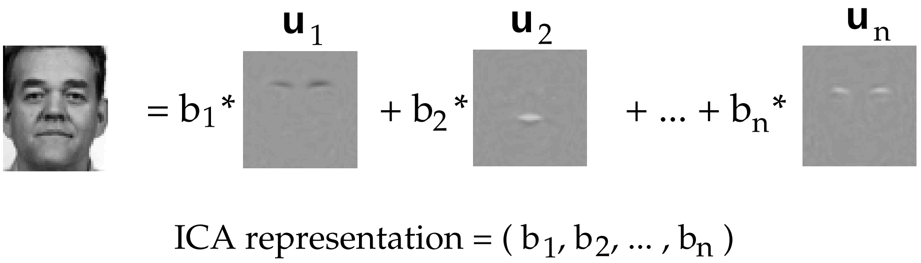

Figure 4: The independent basis image representation consisted of the coefficients, **b**, for the linear combination of independent basis images, **u**, that comprised each face image **x**.

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 $m$ principal component 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, **b**, for the linear combination of basis images in $U$ that comprised the face images in $X$ were determined as follows:

The principal component 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 Equation 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 independent component 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 principal component representation based on the training images to obtain $R_{test} = X_{test} P_m$, and then computing

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

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 $B_{test} = X_{test}(U)^{\dagger}$. $B_{test}$ can be obtained without calculating a pseudo-inverse by normalizing the length of the rows of $U$, thereby making $U$ approximately orthonormal, and calculating $B_{test} = X_{test}$ $U^T$. However, if ICA did not remove all of the second-order dependencies then $U$ will not be precisely orthonormal.

The principal component axes of the Training Set were found by calculating the eigenvectors of the pixelwise covariance matrix over the set of face images. Independent component analysis was then performed on the first $m = 200$ of these eigenvectors, where the first 200 principal components accounted for over 98% of the variance in the images.[4] The $1 \times 3000$ eigenvectors in $P_{200}$ comprised the rows of the $200 \times 3000$ input matrix $X$. The input matrix $X$ was sphered[5] according to Equation 4, and the weights, $W$, were updated according to Equation 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 $U$.

Figure 4 shows a sample of basis images (i.e. rows of $U$) learned in this architecture. 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 non-zero pixels are nearby). This is because a majority of the statistical dependencies are in spatially proximal pixel locations. A set of principal component basis images (PCA axes), are shown in Figure 5 for comparison.

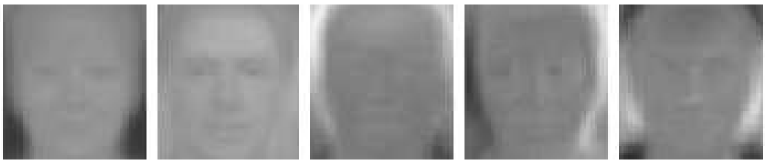

Figure 5: First 5 principal component axes of the image set (columns of $P$).

### 0.4.1 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:

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

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

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.

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

Such normalization is consistent with neural models of primary visual cortex [27]. Cosine similarity measures were previously found to be effective for computational models of language [28] and face processing [55].

Figure 6 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 principal component vectors, which was the eigenface representation used by Pentland, Moghaddam and

[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.

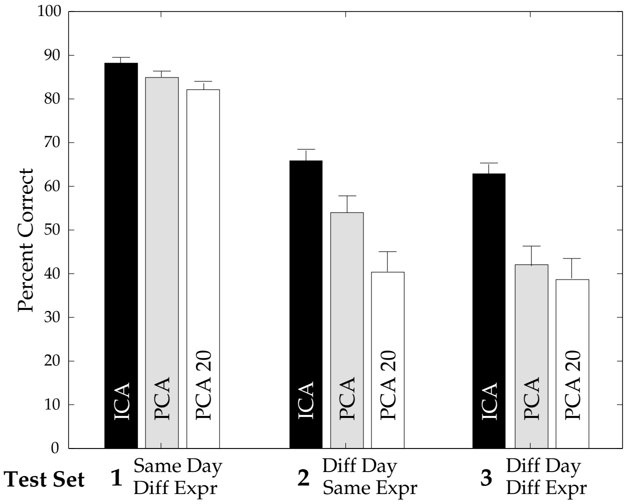

Figure 6: Percent correct face recognition for the ICA representation using 200 independent components, the PCA representation using 200 principal components, and the PCA representation using 20 principal components. Groups are performances for Test Set 1, Test Set 2, and Test Set 3. Error bars are one standard deviation of the estimate of the success rate for a Bernoulli distribution.

Starner [60]. Best performance for PCA was obtained using 200 coefficients. Excluding the first 1, 2, or 3 principal components 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 over all three test sets ($Z = 2.5, p < 0.05$) for Test sets 1 and 2, and ($Z = 2.4, p < 0.05$) for Test Set 3.

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

### 0.4.2 Subspace selection

When all 200 components were retained, then PCA and ICA were working in the same subspace. However, as illustrated in Figure 1, 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:

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

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

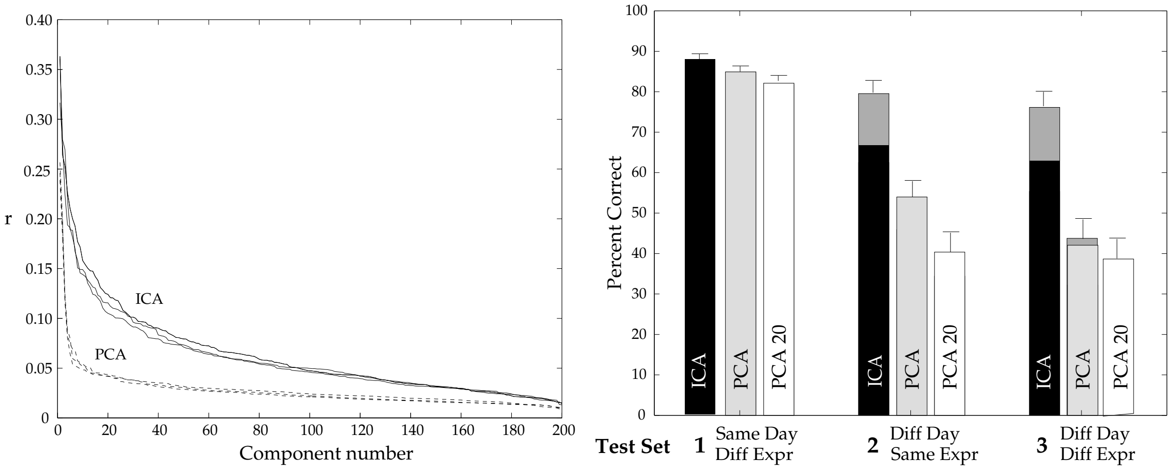

Figure 7: Selection of components by class discriminability. Left: Discriminability of the ICA coefficients (solid lines) and discriminability of the PCA components (dotted lines) for the three test cases. Components were sorted by the magnitude of $r$. Right: Improvement in face recognition performance 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.

of $r$. Figure 7 (Top) compares the discriminability of the ICA coefficients to the PCA coefficients. The ICA coefficients consistently had greater class discriminability than the PCA coefficients.

Face classification performance was compared using the $k$ most discriminable components of each representation. Figure 7 (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 ($Z = 2.9, p < .05; Z = 3.4, p < .01$), respectively. The class discriminability analysis selected subsets of bases from the PCA and ICA representations under the same criterion. Here, the ICA-defined subspace encoded more information about facial identity than PCA-defined subspace.

## 0.5 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 [5, 2]. 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 Figure 2 Right.) This corresponds to treating the columns of $A \triangleq W_I^{-1}$ as a set of basis images. (See Figure **??**.) 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 $X$ (Figure 8). ICA attempts to make the outputs, $U$, as independent as possible. Hence $U$ is a factorial code for the face images.

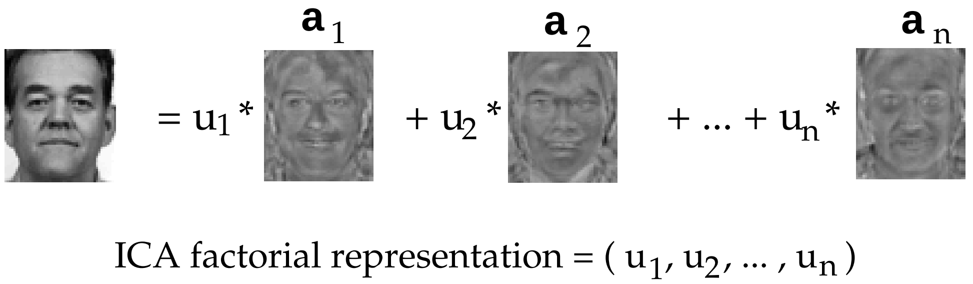

Figure 8: The factorial code representation consisted of the independent coefficients, $\mathbf{u}$, for the linear combination of basis images in $A$ that comprised each face image $\mathbf{x}$.

The representational code for test images is obtained by

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

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 principal components 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:

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

The basis images for this representation consisted of the columns of $A \overset{\Delta}{=} W_I^{-1}$. A sample of the basis images is shown in Figure 8, where the principal component 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 Figure **??**.

### 0.5.1 Face recognition performance: Architecture II

Face recognition performance was again evaluated by the nearest neighbor procedure using cosines as the similarity measure. Figure 9 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.

[6] Here, each pixel has zero mean.

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

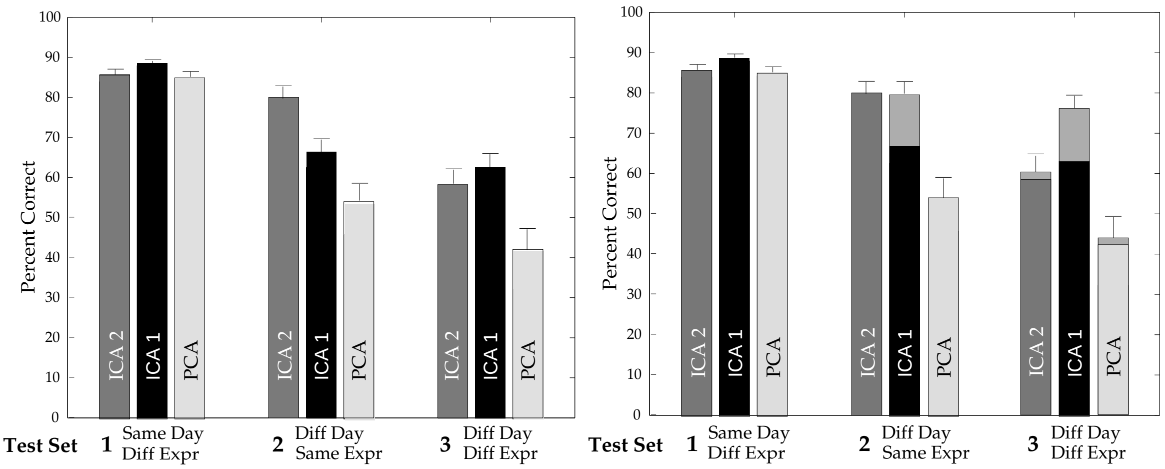

Figure 9: Left: Recognition performance of the factorial code ICA representation (ICA2) using all 200 coefficients, compared to the ICA independent basis representation (ICA1), and the PCA representation, also with 200 coefficients. Right: Improvement in recognition performance of the two ICA representations and the PCA representation by selecting subsets of components by class discriminability. Gray extensions show improvement over recognition performance using all 200 coefficients.

Class discriminability of the 200 ICA factorial coefficients was calculated according to Equation 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 representation by class discriminability had little effect on the recognition performance using the ICA-factorial representation (see Figure 9 right). 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 independent components, 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 independent components, 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 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.

### 0.5.2 Combined ICA recognition system

The similarity spaces in the two ICA representations were not identical. While the the two systems tended to make errors on the same images, they did not assign the same incorrect identity. In [10] we showed that an effective reliability criterion is to accept identifications only when the two systems give the same answer. Under this criterion, accuracy improved to 100%, 100%, and 97% for the three test sets.

Another way to combine the two systems is to define a new similarity measure as the sum of the similarities under Architecture I and Architecture II. In [10] we showed that this combined classifier improved performance to 91.0% ,88.9%, and 81.0% for the three test cases, which significantly outperformed PCA in all three conditions ($Z = 2.7, p < 0.01; Z = 3.7, p < .001; Z = 3.7; p < .001$).

# 0.6 Examination of the ICA Representations

## 0.6.1 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

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

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

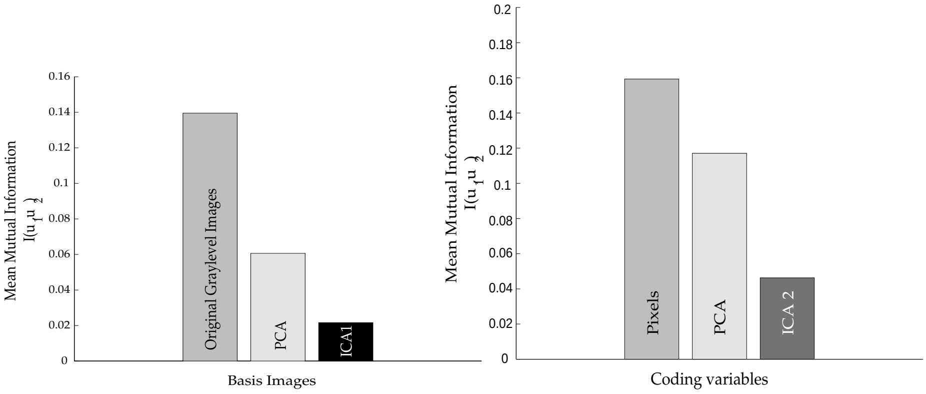

Figure 10: Pairwise mutual information. LEFT: Mean mutual information between basis images. Mutual information was measured between pairs of graylevel images, principal component images, and independent basis images obtained by Architecture I. RIGHT: Mean mutual Information between coding variables. Mutual information was measured between pairs of image pixels in graylevel images, PCA coefficients, and ICA coefficients obtained by Architecture II.

Figure 10 (left) compares the mutual information between *basis images* for the original graylevel images, the principal component basis images, and the ICA basis images obtained in Architecture I. Principal component 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.

Figure 10 (right) compares the mutual information between the *coding variables* in the ICA factorial representation obtained with Architecture II, the PCA representation, and graylevel images. For graylevel images, mutual information was calculated between pairs of pixel locations. For the PCA representation, mutual information was calculated between pairs of principal component 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 principal component representation.

## 0.6.2 Sparseness

Field [22] has argued that sparse distributed representations are advantageous for coding visual stimuli. Sparse representations are characterized by highly kurtotic response distributions, in which

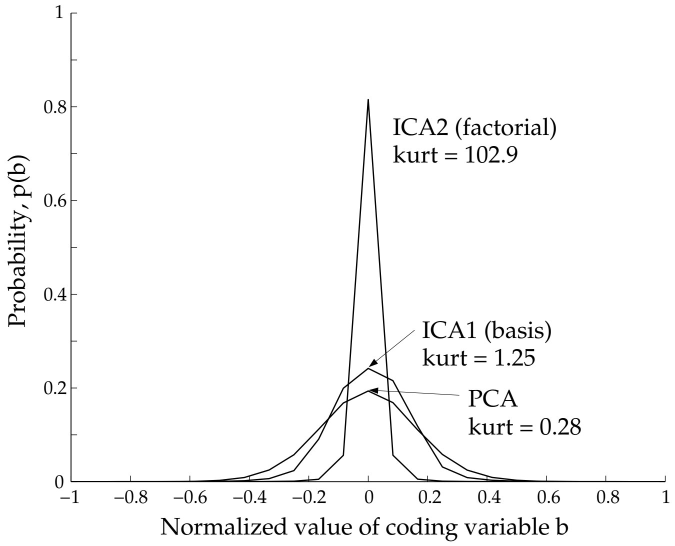

Figure 11: Kurtosis (sparseness) of ICA and PCA representations.

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 [22] 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 contrasts this with a compact code such as principal components, 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 [56, 11].

The probability densities for the values of the coefficients of the two ICA representations and the PCA representation are shown in Figure 11. 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:

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

The kurtosis of the PCA representation was measured for the principal component coefficients. The principal components 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 grayvalues, 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.

[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.

## 0.7 Local basis images versus factorial codes

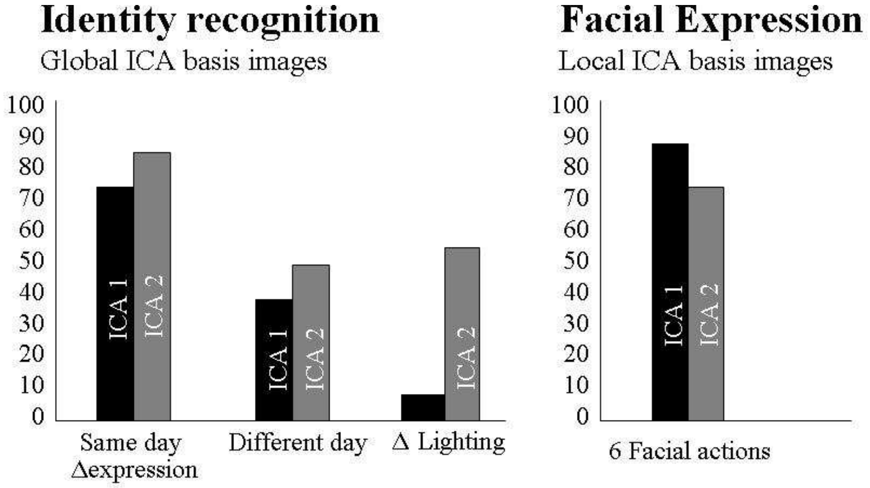

Figure 12: Face recognition performance with a larger image set from Draper et al.(2003). Architecture II outperformed Architecture I for identity recognition, whereas Architecture I outperformed Architecture II for an expression recognition task. Draper and colleagues attributed the findings to local versus global processing demands.

Draper and colleagues [20] conducted a comparison of ICA and PCA on a substantially larger set of FERET face images consisting of 1196 individuals. Performances were compared for L1 and L2 norms as well as cosines and mahalanobis distance. This study supported the findings presented here, namely that ICA outperformed PCA, and this advantage emerged when cosines, but not Euclidean distance, was used as the similarity measure for ICA. This study included a change in lighting condition which we had not previously tested. ICA with architecture II obtained 51% accuracy on 192 probes with changes in lighting, compared to the best PCA performance (with Mahalanobis distance) of 40% correct.

An interesting finding to emerge from the Draper study is that the ICA representation with Architecture II outperformed Architecture I for identity recognition. See Figure 12. According to arguments by Barlow, Atick, and Field [5, 2, 22], the representation defined by Architecture II is a more optimal representation than the Architecture I representation given its sparse, factorial properties. Although Sections 0.4 and 0.5 showed no significant difference in recognition performance for the two architectures, Architecture II had fewer training samples to estimate the same number of free parameters as Architecture I due to the difference in the way the input data was defined. Figure 10 shows that the residual dependencies in the ICA-factorial representation were higher than in the Architecture I representation. In [6] we predicted that The ICA-factorial representation may prove to have a greater advantage given a much larger training set of images. Indeed, this prediction was born out in the Draper study [20].

When the task was changed to recognition of facial expressions, however, Draper et al found that the ICA representation from Architecture I outperformed the ICA representation from Architecture II. The advantage for Architecture I only emerged, however, following subspace selection using the class variability ratios defined in Section 0.4.2. The task was to recognize 6 facial actions, which are individual facial movements approximately corresponding to individual facial muscles. Draper et al attributed their pattern of results to differences in local versus global processing requirements of the two tasks. Architecture I defines local face features whereas Architecture II defines more configural face features. A large body of literature in human face processing points to the importance of configural information for identity recognition, whereas the facial expression recognition task in this study may have greater emphasis on local information. This speaks to the issue of separate basis sets for expression and identity. The neuroscience community has been interested in this distinction, as there is evidence for separate processing of identity and expression in the brain (e.g. [30].) Here we obtain better recognition performance when we define different basis sets for identity versus expression. In the two basis sets we switch what is treated as an observation versus what is treated as an independent variable for the purposes of information maximization.

## 0.8 Discussion

Much of the information that perceptually distinguishes faces may be contained in the higher order statistics of the images. 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 explore one such generalization: Bell and Sejnowski's Infomax 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 & 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 [70], which was based on principal component 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 non-orthogonal, 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. Moreover, with a larger training set, the factorial representation of Architecture II emerged with higher identity recognition performance than Architecture I [20], consistent with theories of the effectiveness of sparse, factorial representations for coding complex visual objects [5, 2, 22].

As discussed in Section 0.2.1, ICA representations are better optimized for transmitting information in the presence of noise than PCA, 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. Draper and colleagues [20], tested a specific form of noise, lighting variation in the FERET dataset, and found a considerable advantage of ICA Architecture II over PCA for robustness to changes in lighting. The Draper study also supported our finding of a substantial advantage of ICA II over PCA for images collected weeks apart. A key challenge in translating any face recognition method into a real-world system is noise. The approach presented here would benefit from a more systematic exploration of tolerance to noise, including the effect of noise at different spatial scales. Moreover it was recently shown that shallower transfer functions than the ones learned by information maximization, proportional to the cube root of the cumulative pdf, optimize information transmission in the presence of noise since error magnification depends on the slope of the transfer function [71]. Thus another avenue of research is to explore face representations based on the optimization function in [71].

A number of research groups have independently tested the ICA representations presented here and in [9, 10]. Liu and Wechsler [42], and Yuen and Lai [76] both supported the finding that ICA outperformed PCA. Moghaddam [48] and Yang [75] employed Euclidean distance as the similarity measure instead of cosines and tested Architecture I only. No differences are expected with Euclidean distance in Architecture I, and consistent with our findings, no significant difference was found. Draper and colleagues [20] conducted a thorough comparison of ICA and PCA using four similarity measures, and supported the findings that ICA outperformed PCA, and this advantage emerged when cosines, but not Euclidean distance, was used as the similarity measure for ICA. Class-specific projections of the ICA face codes using Fisher's linear discriminant has recently been shown to be effective for face recognition as well [32]. ICA was also shown to be effective for facial expression recognition. The ICA representation outperformed more than eight other image representations on a task of facial expression recognition, equaled only by Gabor wavelet decomposition [19, 8], with which it has relationships discussed below.

In this paper, the number of sources was controlled by reducing the dimensionality of the data through principal component analysis prior to performing ICA. There are two limitations to this approach [68]. 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. Secondly, it may not be desirable to throw away subspaces of the data with low power such as the higher principal components. Although low in power, these subspaces may contain independent components, 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 independent components of the data [68]. Techniques for estimating the number of independent components in a dataset have also been proposed [33, 47].

The information maximization algorithm employed to perform independent component analysis in this work assumed that the underlying "causes" of the pixel graylevels 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 [12]. We employed a logistic source model which has shown in practice to be sufficient to separate natural signals with super-Gaussian distributions [12]. The underlying "causes" of the pixel graylevels 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 [37]. 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 [24]. Other influences, such as changes in pose and expression may be linearly approximated only to a limited extent. Nonlinear independent component analysis 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 [38, 74]. 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 [75], and promising results have recently been presented for face recognition with kernel-ICA [43].

Unlike principal component analysis, independent component analysis using Architecture I found a spatially local face representation. Local feature analysis (LFA) [59] also finds local basis images for faces, but using second-order statistics. The LFA basis images are found by performing whitening (Equation 4) on the principal component 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 [19]. Interestingly, downsampling methods based on sequential information maximization significantly improve performance with LFA [58].

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, non-negative matrix factorization (NMF) [35], 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 non-negative weight

[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.

constraints are employed [69]. In experiments to date, RBM's 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 0.2, 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 [39]. 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 in this study.

Desirable filters may be those that are adapted to the patterns of interest and capture interesting structure [40]. 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 [52, 13, 72] and audition [39]. 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.

The principle of independence may have relevance to face and object representations in the brain. Horace Barlow [5] and Joseph 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 [13] 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 [53, 52] obtained a similar result with a sparseness objective, where there is a close information theoretic relationship between sparseness and independence [5, 13]. Conversely, it has also been shown that Gabor filters, which model the responses of V1 simple cells, give outputs that are sparse and independent[10] when convolved with natural scenes but not when convolved with random noise [21, 22, 66]. (See [6] for a discussion).

## 0.9 Face modeling and information maximization: A computational neuroscience perspective

Dependency coding and information maximization appear to be central principles in neural coding early in the visual system. Neural systems with limited dynamic range can increase the information that the response gives about the signal by placing the more steeply sloped portions of the transfer function in the regions of highest density, and shallower slopes at regions of low density. The function that maximizes information transfer is the one that matches the cumulative probability density of the input. There is a large body of evidence that neural codes in vision and other sensory modalities match the statistical structure of the environment, and hence maximize information about environmental signals to a degree. See [67] for a review. These principles may be relevant to how we think about higher visual processes such as face recognition as well.

Here we examine algorithms for face recognition by computer from a perspective of information maximization. Principal component solutions can be learned in neural networks with simple Hebbian learning rules [51]. Hebbian learning is a model for long-term potentiation in neurons, in which weights are increased when the input and output are simultaneously active. The weight update rule is typically formulated as the product of the input and the output activations. This simple rule learns the covariance structure of the input (i.e. the second-order relations). In the case of a single output unit, Hebbian learning maximizes the information transfer between the input and the output [41]. For multiple inputs and outputs, however, Hebbian learning doesn't maximize information transfer

[10] when response normalization is included. There is a large body of evidence for response normalization in visual cortical neurons.

unless all of the signal distributions are Gaussian. In other words, the eigenface representation performs information maximization in the case where the input distributions are Gaussian.

Independent component analysis performs information maximization for a more general set of input distributions. (See [18] for a reference text and [6] for a review). The information maximization learning algorithm employed here was developed from the principle of optimal information transfer in neurons with sigmoidal transfer functions. Inspection of the learning rule in Equation 2 shows that it contains a Hebbian learning term, but it is between the input and the *gradient* of the output. (In the case of the natural gradient learning rule in Equation 3 it is between the input and the natural gradient of the output.) Here we showed that face representations derived from ICA gave better recognition performance than face representations based on PCA. This suggests that information maximization in early processing is an effective strategy for face recognition by computer.

A number of perceptual studies support the relevance of dependency encoding to human face perception. Perceptual effects such as other-race effects are consistent with information maximization coding. For example, face discrimination is superior for same-race than other-race faces [63], which is consistent with a perceptual transfer function that is steeper for faces in the high-density portion of the distribution in an individual's perceptual experience (ie. same-race faces). Face adaptation studies (e.g. [31, 50, 73]) are consistent with information maximization on short time scales. For example, after adapting to a distorted face, a neutral face appears distorted in the opposite direction. Adaptation alters the probability density on short time scales, and the aftereffect is consistent with a perceptual transfer function that has adjusted to match the new cumulative probability density. Adaptation to a nondistorted face does not make distorted faces appear more distorted, which is consistent with an infomax account, because adapting to a neutral face would not shift the population mean. There is also support from neurophysiology for information maximization principles in face coding. The response distributions of IT face cells are sparse and there is very little redundancy between cells [64, 65].

Unsupervised learning of 2nd order dependencies (PCA) successfully models a number of aspects of human face perception including similarity, typicality, recognition accuracy, and other-race effects (e.g. [17, 26, 55]. Moreover, one study found that ICA better accounts for human judgments of facial similarity than PCA, supporting the idea that the more dependencies are encoded, the better the model of human perception for some tasks [25]. The extent to which information maximization principles account for perceptual learning and adaptation aftereffects in human face perception is an open question and an area of active research.

### Acknowledgments

Support for this work was provided by National Science Foundation IIS-0329287, University of California Discovery Program 10158, National Research Service Award MH-12417-02, and Howard Hughes Medical Institute. Portions of the research in this paper use the FERET database of facial images, collected under the FERET program of the Army Research Laboratory. Early versions of this work appear in IEEE Transactions on Neural Networks, 13(6) p. 1450-64, 2002, and in Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Human Vision and Electronic Imaging III, Vol. 3299 p. 528-539, 1998.

# Bibliography

[1] S. Amari, A Cichocki, , and H.H. Yang. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, Cambridge, MA, 1996. MIT Press.

[2] J.J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1992.

[3] J.J. Atick and A.N. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4:196–210, 1992.

[4] F.R. Bach and M.I. Jordan. Kernel independent component analysis. In *Proceedings of the 3rd international conference on independent component analysis and signal separation*, 2001.

[5] H.B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[6] Marian S. Bartlett. *Face Image Analysis by Unsupervised Learning*, volume 612 of *The Kluwer International Series on Engineering and Computer Science*. Kluwer Academic Publishers, Boston, 2001.

[7] M.S. Bartlett. *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. PhD thesis, University of California, San Diego, 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 S.A. Solla, T.K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[9] M.S. Bartlett, H.M. Lades, and T.J. Sejnowski. Independent component representations for face recognition. In T. Rogowitz, B. & Pappas, editor, *Proceedings of the SPIE Symposium on Electonic Imaging: Science and Technology; Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, San Jose, CA, 1998. SPIE Press.

[10] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski. Image representations for facial expression recognition. *IEEE transactions on neural networks*, 13(6):1450–1464, 2002.

[11] E.B. Baum, J. Moody, and F. Wilczek. Internal representaions for associative memory. *Biological Cybernetics*, 59:217–228, 1988.

[12] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[13] A.J. Bell and T.J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[14] A. Cichocki, R. Unbehauen, and E. Rummert. Robust learning algorithm for blind separation of signals. *Electronics Letters*, 30(7):1386–1387, 1994.

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

[16] G. Cottrell and J. Metcalfe. Face, gender and emotion recognition using holons. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 3, pages 564–571, San Mateo, CA, 1991. Morgan Kaufmann.

[17] GW Cottrell, MN Dailey, C Padgett, and Adolphs R. *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, chapter Is all face processing holistic? The view from UCSD. Erlbaum, 2000.

[18] T. M. Cover and J. A. Thomas. *Elements of information theory.* John Wiley & Sons, New York, 1991.

[19] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):974–989, 1999.

[20] B.A. Draper, K. Baek, Bartlett M.S., and J.R. Beveridge. Recognizing faces with pca and ica. *Computer Vision and Image Understanding, Special Issue on Face Recognition*, 91:115–137, 2003.

[21] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America, A*, 4:2379–94, 1987.

[22] D.J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[23] M. Girolami. *Advances in Independent Component Analysis.* Springer-Verlag, Berlin, 2000.

[24] P. Hallinan. *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions.* PhD thesis, Harvard University, 1995.

[25] P. Hancock. Alternative representations for faces. In *British Psychological Society, Cognitive Section.* University of Essex, September 6-8, 2000.

[26] P.J.B. Hancock, A.M. Burton, and V. Bruce. Face processing: human perception and principal components analysis. *Memory and Cognition*, 24:26–40, 1996.

[27] D.J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–197, 1992.

[28] G. Hinton and T. Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

[29] C. Jutten and J. Herault. Blind separation of sources i. an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

[30] Haxby JV, Hoffman EA, and Gobbini MI. The distributed human neural system for face perception. *Trends in Cognitive Science*, 4:223–233, 2000.

[31] D. Kaping, P. Duhamel, and M. Webster. Adaptation to natural face categories. In *Journal of Vision*, volume 2, page 128, 2002.

[32] J. Kim, J. Choi, and J. Yi. Face recognition based on ica combined with fld. In *European Conference on Computer Vision*, pages 10–18, 2002.

[33] H. Lappalainen and J. W. Miskin. Ensemble learning. In M. Girolami, editor, *Advances in Independent Component Analysis*, pages 76–92. Springer-Verlag, 2000.

[34] S. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch*, 36:910–912, 1981.

[35] D.D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[36] T.-W. Lee. *Independent Component Analysis: Theory and Applications.* Kluwer Academic Publishers, 1998.

[37] 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 Computation*, 11(2):417–41, 1999.

[38] T-W. Lee, B.U. Koehler, and R. Orglmeister. Blind source separation of nonlinear mixing models. In *Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing*, pages 406–415, Florida, September 1997.

[39] M. Lewicki and B. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16(7):1587–601, 1999.

[40] M. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–65, 2000.

[41] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[42] C. Liu and H. Wechsler. Comparative assessment of independent component analysis (ica) for face recognition. In *International conference on audio and video based biometric person authentication*, 1999.

[43] Q. Liu, J. Cheng, H. Lu, and S. Ma. Modeling face appearance with nonlinear independent component analysis. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.

[44] D. J. C. MacKay. Maximum likelihood and covariant algorithms for independent component analysis. in preparation, 1996.

[45] S. Makeig, A.J. Bell, T-P. Jung, and T.J. Sejnowski. Independent component analysis of electroencephalographic data. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 145–151, Cambridge, MA, 1996. MIT Press.

[46] 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*, 6(3):160–88, 1998.

[47] J. W. Miskin and D. J. C. MacKay. *Ensemble Learning for Blind Source Separation ICA: Principles and Practice.* Cambridge University Press, 2001. In press.

[48] B. Moghaddam. Principal manifolds and bayesian subspaces for visual recognition. In *International conference on computer vision*, 1999.

[49] J-P. Nadal and N. Parga. Non-linear neurons in the low noise limit: a factorial code maximizes information trans fer. *Network*, 5:565–581, 1994.

[50] M. Ng, D. Kaping, M.A. Webster, S. Anstis, and I. Fine. Selective tuning of face perception. In *Journal of Vision*, volume 3, page 106a, 2003.

[51] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

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

[53] B.A. Olshausen and D.J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333–340, 1996.

[54] A.V. Oppenheim and J.S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69:529–541, 1981.

[55] A. O'Toole, K. Deffenbacher, D. Valentin, and H. Abdi. Structural aspects of face recognition and the other race effect. *Memory and Cognition*, 22(2):208–224, 1994.

[56] G. Palm. On associative memory. *Biological Cybernetics*, 36:19–31, 1980.

[57] Barak A. Pearlmutter and Lucas C. Parra. A context-sensitive generalization of ica. In Mozer, Jordan, and Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press,, 1996.

[58] P.S. Penev. Redundancy and dimensionality reduction in sparse-distributed representations of natural objects in terms of their local features. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[59] P.S. Penev and J.J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7(3):477–500, 1996.

[60] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[61] P.J. Phillips, H. Wechsler, J. Juang, and P.J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306, 1998.

[62] L.N. Piotrowski and F.W. Campbell. A demonstration of the visual importance and flexibility of spatial-frequency, amplitude, and phase. *Perception*, 11:337–346, 1982.

[63] Walker PM and Tanaka JW. An encoding advantage for own-race versus other-race faces. *Perception*, 32(9):1117–25, 2003.

[64] E.T. Rolls, N.C. Aggelopoulos, L. Franco, and A. Treves. Information encoding in the inferior temporal cortex: contributions of the firing rates and correlations between the firing of neurons. *Biological Cybernetics*, 90:19–32, 2004.

[65] E.T. Rolls and M.J. Tovee. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology*, 73(2):713–726, 1995.

[66] E. P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *31st Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2-5 1997.

[67] E.O. Simoncelli. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–1216, 2001.

[68] J.V. Stone and J. Porrill. Undercomplete independent component analysis for signal separation and dimension reduction. Technical report, University of Sheffield, Department of Psychology, 1998.

[69] Y.W. Teh and G.E. Hinton. Rate-coded restricted boltzmann machines for face recognition. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[70] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[71] T. von der Twer and D.I.A. Macleod. Optimal nonlinear codes for the perception of natural colors. *Network: Computation in Neural Systems*, 12:395–407, 2001.

[72] T. Wachtler, T.-W. Lee, and T.J. Sejnowski. The chromatic structure of natural scenes. *Journal of the Optical Socitey of America, A*, 18(1):65–77, 2001.

[73] M.M. Webster. Figural aftereffects in the perception of faces. *Psychonomic Bulletin Review*, 6(4):647–653, 1999.

[74] H.H. Yang, S.-I. Amari, and A. Cichocki. nformation-theoretic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–3000, 1998.

[75] M. Yang. Face recognition using kernel methods. In T. Diederich, Becker S., and Z. Ghahramani, editors, *Advances in neural information processing systems*, volume 14, 2002.

[76] P.C. Yuen and J.H. Lai. Independent component analysis of face images. In *IEEE workshop on biologically motivated computer vision*, Seoul, 2000. Springer-Verlag.