Independent component representations for face recognition*

Marian Stewart Bartlett[a], H. Martin Lades[b], and Terrence J. Sejnowski[c]

[a]University of California San Diego and the Salk Institute, La Jolla, CA 92037.

[b]Lawrence Livermore National Laboratories, Livermore, CA 94550.

[c]University of California San Diego and Howard Hughes Medical Institute at the Salk Institute, La Jolla, CA 92037.

ABSTRACT

In a task such as face recognition, much of the important information may be contained in the high-order relationships among the image pixels. A number of face recognition algorithms employ principal component analysis (PCA), which is based on the second-order statistics of the image set, and does not address high-order statistical dependencies such as the relationships among three or more pixels. Independent component analysis (ICA) is a generalization of PCA which separates the high-order moments of the input in addition to the second-order moments. ICA was performed on a set of face images by an unsupervised learning algorithm derived from the principle of optimal information transfer through sigmoidal neurons.[1] The algorithm maximizes the mutual information between the input and the output, which produces statistically independent outputs under certain conditions. ICA was performed on the face images under two different architectures. The first architecture provided a statistically independent basis set for the face images that can be viewed as a set of independent facial features. The second architecture provided a factorial code, in which the probability of any combination of features can be obtained from the product of their individual probabilities. Both ICA representations were superior to representations based on principal components analysis for recognizing faces across sessions and changes in expression.

Keywords: Independent component analysis, ICA, principal component analysis, PCA, face recognition.

1. INTRODUCTION

Several advances in face recognition such as "Holons,[2] " "Eigenfaces,[3] " and "Local Feature Analysis[4]" have employed forms of principal component analysis, which addresses only second-order moments of the input. Principal component analysis is optimal for finding a reduced representation that minimizes the reconstruction error, but the axes that account for the most reconstruction error may not be optimal for coding aspects of the image relevant to classification.[5] Independent component analysis (ICA) is a generalization of principal component analysis (PCA), which decorrelates the high-order moments of the input in addition to the second order moments.[6] In a task such as face recognition, much of the important information is contained in the high-order statistics of the images. A representational basis in which the high-order statistics are decorrelated may be more powerful for face recognition than one in which only the second-order statistics are decorrelated, as in PCA representations.

Bell and Sejnowski[1] recently developed an algorithm for separating the statistically independent components of a dataset through unsupervised learning. This algorithm has proven successful for separating randomly mixed auditory signals (the cocktail party problem), and has recently been applied to separating EEG signals,[7] fMRI signals,[8] and finding image filters that give independent outputs from natural scenes.[9]

We developed methods for representing face images for face recognition based on ICA using two architectures. The first architecture corresponded to that used to perform blind separation of a mixture auditory signals[1] and of EEG[7] and fMRI data.[8] We employed this architecture to find a set of statistically independent source images for a set of face images. These source images comprised a set of independent basis images for the faces, and can be viewed as set of statistically independent image features, in which the pixel values in one image feature cannot be predicted from the pixel values of the other image features.

---

*In Press: *Proceedings of the SPIE Symposium on Electronic Imaging: Science and Technology; Conference on Human Vision and Electronic Imaging III*, San Jose, California, January, 1998.

Email: marni@salk.edu(M.S. Bartlett),terry@salk.edu (T.J. Sejnowski). lades@earthlink.net (H.M. Lades).
Website: http//www.cnl.salk.edu/ marni

The second architecture corresponded to that used to find image filters that produced statistically independent outputs from natural scenes.[9] We employed this architecture to define a set of statistically independent variables for representing face images. In other words, we used ICA under this architecture to find a factorial code for the face 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.[10,11]

Face recognition performance was tested using the FERET database.[12] Face recognition performances using the ICA representations were benchmarked by comparing them to recognition performances using the "Eigenface" representation, which is based on PCA.[3,13]

2. INDEPENDENT COMPONENT ANALYSIS (ICA)

ICA is an unsupervised learning rule that was derived from the principle of optimal information transfer through sigmoidal neurons.[14,1] Consider the case of a single input, x, and output, y, passed through a nonlinear squashing function, g, as illustrated in Figure 1.

$$u = wx + w_0 \qquad y = g(u) = \frac{1}{1+e^{-u}}$$

The optimal weight w on x for maximizing information transfer is the one that best matches the probability density of x to the slope of the nonlinearity. The optimal w produces the flattest possible output density, which in other words, maximizes the entropy of the output.

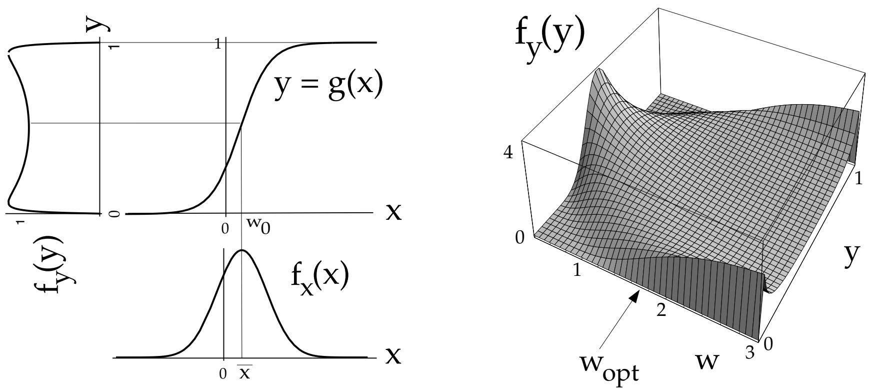

Figure 1. Optimal information flow in sigmoidal neurons. The input x is passed through a nonlinear function, $g(x)$. The information in the output density $f_y(y)$ depends on matching the mean and variance of $f_x(x)$ to the slope and threshold of $g(x)$. Right: $f_y(y)$ is plotted for different values of the weight, w. The optimal weight, w_{opt} transmits the most information. Figure from Bell & Sejnowski (1995), reprinted with permission from Neural Computation.

The optimal weight is found by gradient ascent on the entropy of the output, y with respect to w. When there are multiple inputs and outputs, maximizing the joint entropy of the output encourages the individual outputs to move towards statistical independence. When the form of the nonlinear transfer function g is the same as the cumulative density functions of the underlying independent components (up to a scaling and translation) it can be shown that maximizing the mutual information between the input X and the output Y also minimizes the mutual information between the u_i.[15,9]

The update rule for the weight matrix, W, for multiple inputs and outputs is given by

$$\Delta W = (I + y' u^T) W \tag{1}$$
$$\text{where } y' = \frac{\partial}{\partial y_i}\frac{\partial y_i}{\partial u_i} = \frac{\partial}{\partial u_i} ln \frac{\partial y_i}{\partial u_i}.$$

We employed the logistic transfer function, $g(u) = \frac{1}{1+e^{-u}}$, giving $y' = (1 - 2y_i)$.

The algorithm includes a "sphering" step prior to learning.[9] The row means are subtracted from the dataset, X, and then X is passed through the zero-phase whitening filter, W_z, which is twice the inverse square root of the covariance matrix:

$$W_z = 2 * \langle XX^T \rangle^{-\frac{1}{2}}. \tag{2}$$

This removes both the first and the second order statistics of the data; both the mean and covariances are set to zero and the variances are equalized. The full transform from the zero-mean input was calculated as the product of the sphering matrix and the learned matrix, $W_I = W * W_Z$. The pre-whitening filter in the ICA algorithm has the Mexican-hat shape of retinal ganglion cell receptive fields which remove much of the variability due to lighting.[9]

3. INDEPENDENT COMPONENT REPRESENTATIONS OF FACE IMAGES

3.1. Statistically independent basis images

To find a set of statistically independent basis images for the set of faces, we separated the independent components of the face images according to the image synthesis model of Figure 2. The face images in X were assumed to be a linear mixture of an unknown set of statistically independent source images S, where A is an unknown mixing matrix. The sources were recovered by a matrix of learned filters, W_I, which produced statistically independent outputs, U. This synthesis model is related to that used to perform blind separation on an unknown mixture of auditory signals[1] and to separate the sources of EEG signals[7] and fMRI images.[8]

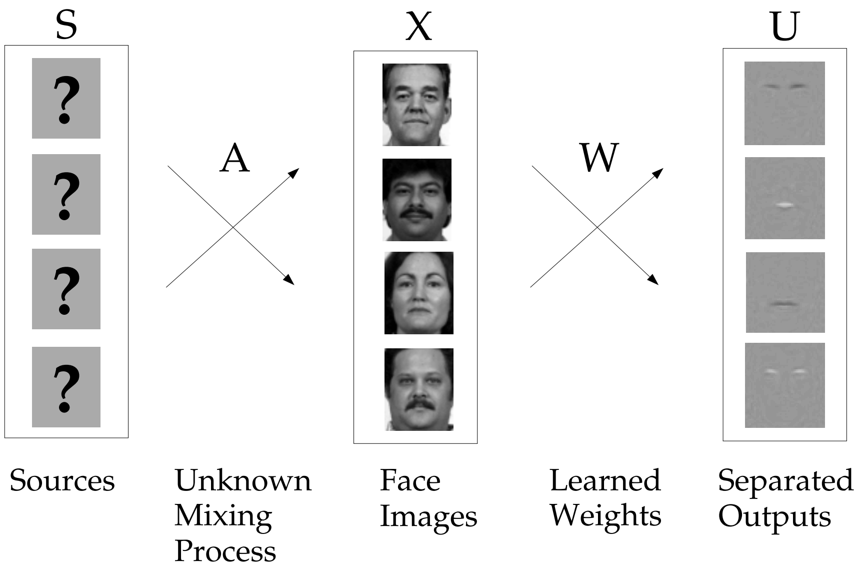

Figure 2. Image synthesis model. For finding a set of independent component images, the images in X are considered to be a linear combination of statistically independent basis images, S, where A is an unknown mixing matrix. The basis images were recovered by a matrix of learned filters, W_I, that produced statistically independent outputs, U.

The images comprised the rows of the input matrix, X. With the input images in the rows of X, the ICA outputs in the rows of $W_I X = U$ were also images, and provided a set of independent basis images for the faces (Figure 3). These basis images can be considered a set of statistically independent facial features, where the pixel values in each feature image were statistically independent from the pixel values in the other feature images. The ICA representation consisted of the coefficients for the linear combination of independent basis images in U that comprised each face image (Figure 3).

The number of independent components found by the ICA algorithm corresponds with the dimensionality of the input. In order to have control over the number of independent components extracted by the algorithm, instead of performing ICA on the n original images, we performed ICA on a set of m linear combinations of those images,

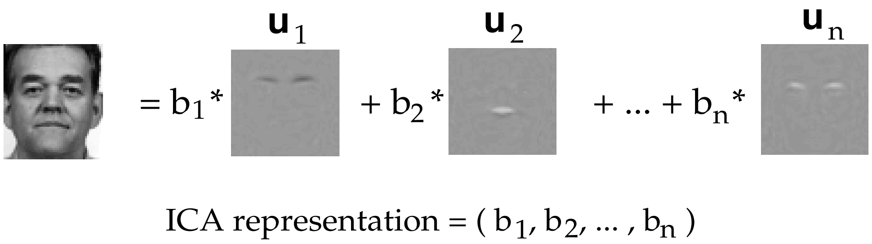

Figure 3. The independent basis image representation consisted of the coefficients, **b**, for the linear combination of independent basis images, **u**, that comprised each face image **x**.

where $m < n$. 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,[8] we chose for these linear combinations the first m principal component vectors of the image set. PCA gives the linear combination of the parameters (in this case, 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 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. 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 $R_m = X * P_m$. A minimum squared error approximation of X is obtained by $X_{rec} = R_m * P_m^T$.

The ICA algorithm produced a matrix $W_I = W * W_Z$ such that

$$W_I * P_m^T = U \qquad \Rightarrow \qquad P_m^T = W_I^{-1} U.$$

Therefore

$$X_{rec} = R_m * P_m^T \qquad \Rightarrow \qquad X_{rec} = R_m * W_I^{-1} U.$$

Hence the rows of $R_m * W_I^{-1}$ contained the coefficients for the linear combination of statistically independent sources U that comprised X_{rec}, where X_{rec} 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

$$B = R_m * W_I^{-1}. \tag{3}$$

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 $B_{test} = R_{test} * W_I^{-1}$.

3.2. A factorial code

The previous analysis produced statistically independent basis images. The representational code consisted of the set of coefficients for the linear combination of the independent basis images from which each face image could be reconstructed. Although the basis images were spatially independent, the coefficients were not. By altering the architecture of the independent component analysis, we defined a second representation in which the *coefficients* were statistically independent, in other words, the new ICA outputs formed a factorial code for the face images. Instead of separating the face images to find sets of independent *images*, as in Architecture 1, we separated the *elements* of the face representation to find a set of independent variables for coding the faces. The alteration in architecture corresponded to transposing the input matrix X such that the images were in columns and the pixels in rows (see

Figure 4). Under this architecture, the filters (rows of W_I) were images, as were the columns of $A = W_I^{-1}$. The columns of A formed a new set of basis images for the faces, and the coefficients for reconstructing each face were contained in the columns of the ICA outputs, U.

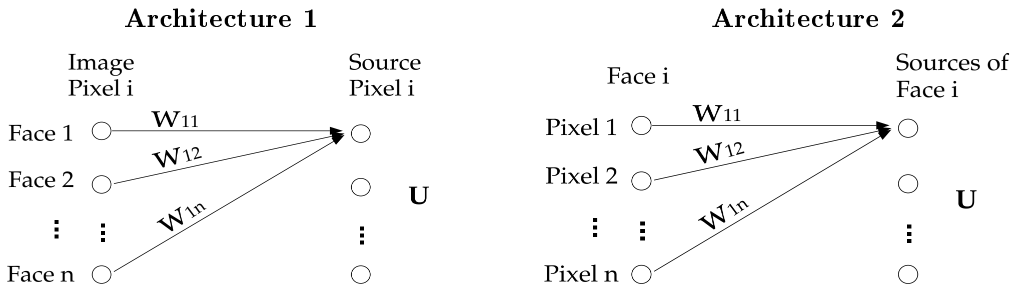

Figure 4. Two architectures for performing ICA on images. Left: Architecture for finding statistically independent basis images. Performing source separation on the face images produced independent component images in the rows of U. Right: Architecture for finding a factorial code. Performing source separation on the pixels produced a factorial code in the columns of the output matrix, U.

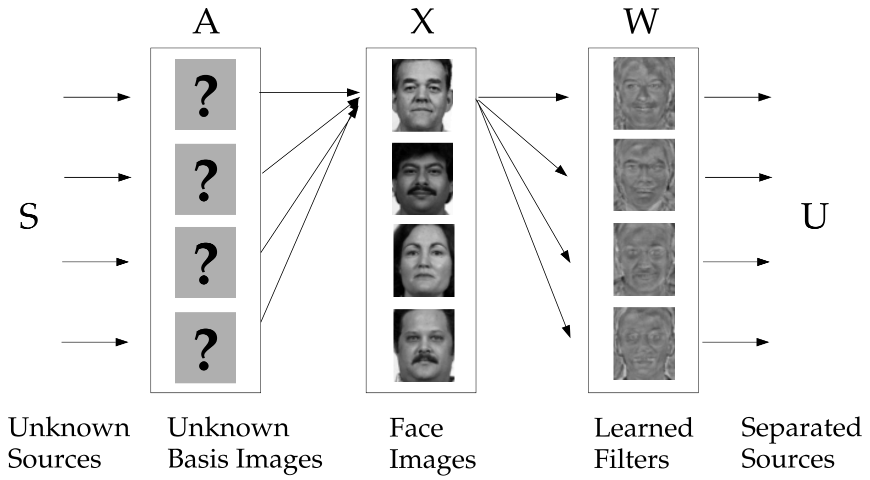

Figure 5. Image synthesis model for Architecture 2, based on Olshausen & Field (1996) and Bell & Sejnowski (1997). Each image in the dataset was considered to be a linear combination of underlying basis images, given by the matrix A. The basis images were each associated with a set of independent "causes", given by a vector of coefficients in S. The causes were recovered by a matrix of learned filters, W_I, which attempts to invert the unknown basis functions to produce statistically independent outputs, U.

Architecture 2 is associated with the image synthesis model of Olshausen and Field,[16] and was also employed by Bell and Sejnowski[9] for finding image filters that produced statistically independent outputs from natural scenes. (See Figure 6.) Images were considered to be created from a set of basis images in A and a vector of underlying statistically independent image causes, in S. The ICA algorithm attempts to invert the basis images by finding a set of filters W_I that produce statistically independent outputs. This image synthesis model differs from that in Figure 2 in that the basis images are the columns of $A = W_I^{-1}$, and the statistically independent sources, U, are the coefficients.

The columns of the ICA output matrix, $W_I X = U$, provided a factorial code for the training images in X. Each column of U contained the coefficients of the the basis images in A for reconstructing each image in X (Figure 6). The representational code for test images was found by $W_I X_{test_i} = U_{test_i}$, where X_{test} was the zero-mean matrix of test images, and W_I was the weight matrix found by performing ICA on the training images.

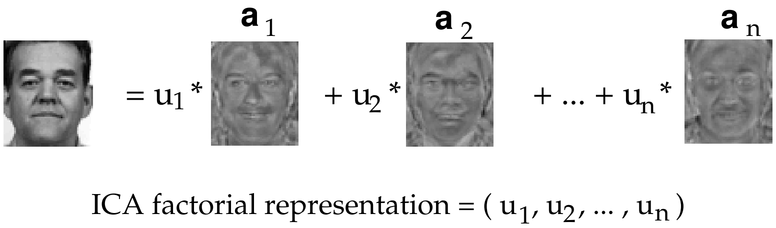

Figure 6. The factorial code representation consisted of the independent coefficients, $\mathbf{u}$, for the linear combination of basis images in A that comprised each face image $\mathbf{x}$.

4. FACE RECOGNITION PERFORMANCE

Face recognition performance was evaluated for the two ICA representations using the FERET face database.[12] 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 Figure 7. The two algorithms were trained on a single frontal view of each individual, and tested for recognition under three different conditions: same session, different expression; different session, same expression; and different session, different expression (see Table 1).

Figure 7. 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.

Image Set	**Condition**		**Number of Images**
Training Set	Session II	50% neutral 50% other	425
Test Set 1	Same Session	Different Expression	424
Test Set 2	Different Session	Same Expression	60
Test Set 3	Different Session	Different Expression	59

Table 1. Image sets used for training and testing.

Coordinates for eye and mouth locations were provided with the FERET database. These coordinates were used to center the face images, crop and scale them to 60×50 pixels based on the area of the triangle defined by the eyes and mouth. The luminance was normalized. For the subsequent analyses, the rows of the images were concatenated to produce 1×3000 dimensional vectors.

4.1. Independent basis architecture

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 200 of these eigenvectors, P_{200}. The 1×3000 eigenvectors in P_{200} comprised the rows of the 200×3000 input matrix X. The input matrix X was sphered according to Equation 2, and the weights, W, were updated according to Equation 1 for 1600 iterations. The learning rate was initialized at 0.001 and annealed down to 0.0001. Training

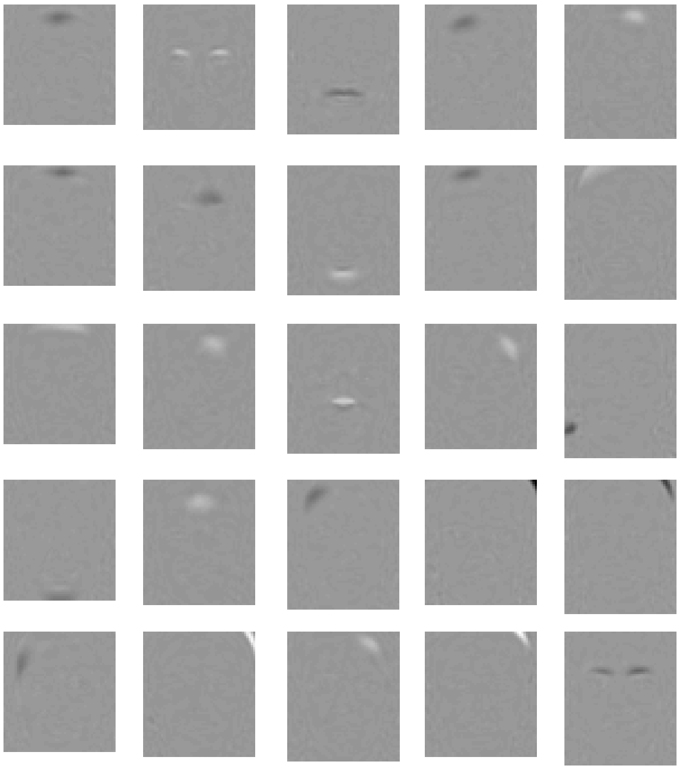

Figure 8. Twenty-five independent components of the image set, which provide a set of statistically independent basis images. Independent components are ordered by the class discriminability ratio, r.

took 90 minutes on a Dec Alpha 2100a quad processor. Following training, a set of statistically independent source images were contained in the rows of the output matrix U.

Figure 8 shows a subset of 25 source images. A set of principal component basis images (PCA axes), are shown in Figure 9 for comparison. The ICA basis images were more spatially local than the principal component basis images. Two factors contribute to the local property of the ICA basis images: The ICA algorithm produces sparse outputs,[9] and secondly, most of the statistical dependencies may be in spatially proximal image locations.

These source images in the rows of U were used as the basis of the ICA representation. The coefficients for the zero-mean training images were contained in the rows of $B = R_{200} * W_I^{-1}$ according to Equation 3, and coefficients for the test images were contained in the rows of $B_{test} = R_{Test} * W_I^{-1}$ where $R_{Test} = Test * P_{200}$.

Face recognition performance was evaluated for the coefficient vectors $\mathbf{b}$ by the nearest neighbor algorithm. Coefficient vectors in the test set were assigned the class label of the coefficient vector in the training set with the most similar angle, as evaluated by the cosine:

$$d = \frac{\mathbf{b}_{test} \cdot \mathbf{b}_{train}}{\| \mathbf{b}_{test} \| \| \mathbf{b}_{train} \|}. \tag{4}$$

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. Figure 10 gives face recognition performance with both the ICA and the PCA based representations. Face recognition performance with the ICA representation was superior to that with the PCA representation. Recognition performance is also shown for the PCA based representation using the first 20 principal component vectors, which was the representation used by Pentland, Moghaddam and Starner.[13] 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.

Face recognition performances for the PCA and ICA representations were next compared by selecting subsets of components by class discriminability. Let $\overline{x}$ be the overall mean of a coefficient, and $\overline{x}_j$ be the mean for person j.

Figure 9. First 25 principal components of the image set, ordered left to right, top to bottom, by the magnitude of the corresponding eigenvalue.

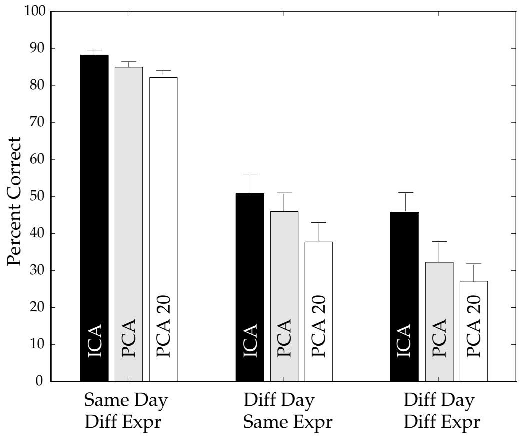

Figure 10. Percent correct face recognition for the ICA representation using 200 independent components, the PCA representation using 200 principal components, and the PCA representation using 20 principal components. Groups are performances for Test Set 1, Test Set 2, and Test Set 3. Error bars are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

For both the PCA and ICA representations, we calculated the ratio of between-class to within-class variability, r, for each coefficient.

$$r = \frac{\sigma_{between}}{\sigma_{within}} \tag{5}$$

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.

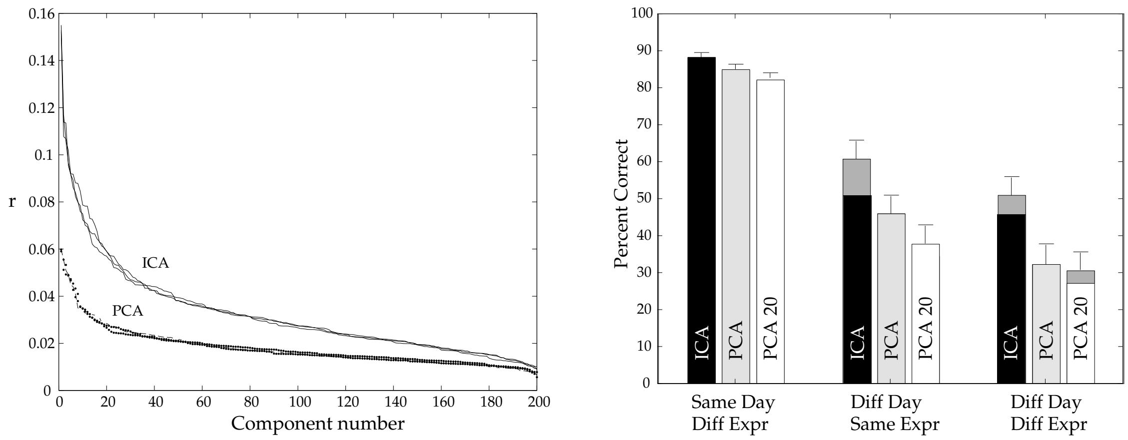

Figure 11. Left: Discriminability of the ICA coefficients (solid lines) and discriminability of the PCA components (dotted lines) for the three test cases. Components were sorted by the magnitude of r. Right: Percent correct face recognition for the ICA and PCA representations using subsets of components selected by the class discriminability r. The improvement is indicated by the gray segments at the top of the bars.

The class discriminability analysis was carried out using the 59 subjects for which four frontal view images were available. The ratios r were calculated separately for each test set, excluding the images from the corresponding test set from the analysis. Both the PCA and ICA coefficients were then ordered by the magnitude of r. Figure 11 (left) 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 11 (right) shows the best classification performance obtained for the PCA and ICA representations, which was with the 100 most discriminable components for both representations. 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, and the advantage of ICA increased as the difference between viewing conditions increased.

4.2. Factorial code architecture

ICA was again performed on the face images using Architecture 2. Instead of performing ICA directly on the 3000 image pixels, ICA was performed on the first 200 PCA coefficients of the face images in order to reduce the dimensionality. The first 200 principal components accounted for over 98% of the variance in the images. These coefficients comprised the columns of the input data matrix, $X = R_{200}^T$.

The ICA algorithm found a 200 $\times$ 200 weight matrix W_I that produced a set of independent coefficients in the output. The basis functions for this representation consisted of the columns of $A = W_I^{-1}$. A sample of the basis set is shown in Figure 12, where the principal component reconstruction $P_{200}A$ was used to visualize the bases as images. The basis images in A have more global properties than the basis images in ICA output of Architecture 1 (Figure 8). Unlike the ICA output, U, nothing forces the columns of A to be sparse.

The columns of U contained the representational codes for the training images. The representational code for the test images was found by $W_I X_{test} = U_{test}$, where X_{test} was the zero-mean matrix of the test images. This produced 200 coefficients for each face image, consisting of the outputs of the 200 ICA filters.

Face recognition performance was evaluated by the nearest neighbor procedure of Section 3.1. Figure 13 compares the face recognition performance using the ICA factorial code representation to the ICA representation using independent basis images and to the PCA representation, each with 200 coefficients. The two ICA representations gave similar recognition performance, and both outperformed the PCA representation. Class discriminability of the new ICA coefficients was calculated according to Equation 5. Again, the ICA coefficients had consistently higher class discriminabilility than the PCA coefficients. Face recognition performance was reevaluated using subsets of components selected by class discriminability. Best performance was obtained with the 180 most discriminable components. The improvement in face recognition obtained by the discriminability analysis is illustrated by the gray extensions in Figure 13.

Figure 12. Basis images for the ICA factorial representation, obtained by training on the first 200 principal component *coefficients* of the face images, oriented in columns of the input. Bases were contained in the columns of $A = W_I^{-1}$.

5. DISCUSSION

In a task such as face recognition, much of the important information may be contained in the high-order spatial relationships among the pixels. Statistically independent representations proved to be more powerful for face recognition than a principal component based representation such as "Eigenfaces," in which only the second-order statistics are decorrelated. Two forms of independent component representations for face images were derived. The first representation was based on a set of statistically independent basis images, which can be viewed as a set of statistically independent facial features. The second representation was based on a factorial code, in which the coefficients were statistically independent. Face recognition performance with the ICA based representations was compared to that with principal component based representations. The analysis was performed on the FERET face database.[12]

Both ICA representations outperformed the "Eigenface" representation, which was based on principal components, for recognizing images of faces sampled on a different day from the training images. The advantage of ICA increased as the difference in imaging conditions increased. This result is particularly encouraging, since the task of most face recognition applications is to identify images collected on a different day from the sample images. These images can differ in the precise lighting conditions and camera angle, and there can be gross differences due to changes in hair, make-up, and facial expression.

The ICA representation of faces presented in Section 3.1 is related to the method of local feature analysis (LFA), which is a topographic representation based on principal components analysis. The LFA kernels are found by performing zero phase whitening (Equation 2) on the principal component axes, followed by a rotation to topographic correspondence with pixel location. In LFA, the kernels are optimally matched to the second-order statistics of the input ensemble, whereas in the ICA representation, the kernels are optimally matched to the high-order statistics of the ensemble as well as the second-order statistics. Both methods produce local filters, although no explicit locality constraint was built into either algorithm.

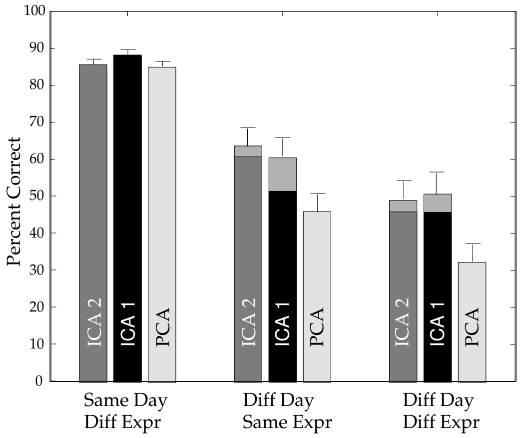

Figure 13. Recognition performance of the factorial code ICA representation (ICA 2) using all 200 coefficients, compared to the ICA independent basis representation (ICA 1), and the PCA representation, also with 200 coefficients. Gray extensions show improvement in recognition performances using a subset of coefficients with the highest class discriminability.

In Section 3.2, independent component analysis 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.[10,11] The two ICA representations were equally powerful for face recognition and both outperformed a representation based on only the second-order relationships in the face images.

ACKNOWLEDGMENTS

We are grateful to Martin McKeown and Michael Gray for helpful discussions on this topic. Support for this work was provided by Lawrence Livermore National Laboratories ISCR agreement B291528, the McDonnell-Pew Center for Cognitive Neuroscience at San Diego, and the Howard Hughes Medical Institute.

REFERENCES

1. A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation* **7**, pp. 1129–1159, 1995.
2. G. Cottrell and J. . Metcalfe, "Face, gender and emotion recognition using holons," in *Advances in Neural Information Processing Systems*, D. Touretzky, ed., vol. 3, pp. 564–571, Morgan Kaufman, (San Mateo, CA), 1991.
3. M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cog. Neurosci.* **3**(1), pp. 71–86, 1991.
4. P. Penev and J. Atick, "Local feature analysis: a general statistical theory for object representation," *Network: Computation in Neural Systems* **7**(3), pp. 477–500, 1996.
5. H. O'Toole, Abdi, K. Deffenbacher, and D. Velantin, "Low-dimensional representation of faces in higher dimensions of the face space.," *Journal of the Optical Society of America A* **10**(3), pp. 405–411, 1993.
6. P. Comon, "Independent component analysis - a new concept?," *Signal Processing* **36**, pp. 287–314, 1994.
7. S. Makeig, A. Bell, T.-P. Jung, and T. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems*, D. Touretzky, M. Mozer, and M. Hasselmo, eds., vol. 8, pp. 145–151, MIT Press, (Cambridge, MA), 1996.
8. M. McKeown, S. Makeig, G. Brown, T.-P. Jung, S. Kindermann, A. Bell, and T. Sejnowski, "Analysis of fmri data by decomposition into independent components," *Proc. Nat. Acad. Sci.* , in press.
9. A. Bell and T. Sejnowski, "The independent components of natural scenes are edge filters," *Vision Research* **37**(23), pp. 3327–3338, 1997.
10. H. Barlow, "Unsupervised learning," *Neural Computation* **1**, pp. 295–311, 1989.

11. J. Atick, “Could information theory provide an ecological theory of sensory processing?,” *Network* **3**, pp. 213–251, 1992.
12. P. Phillips, H. Wechsler, J. Juang, and P. Rauss, “The feret database and evaluation procedure for face-recognition algorithms,” *Image and Vision Computing Journal* , in press.
13. 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.
14. S. Laughlin, “A simple coding procedure enhances a neuron's information capacity,” *Z. Naturforsch* **36**, pp. 910–912, 1981.
15. J.-P. Nadal and N. Parga, “Non-linear neurons in the low noise limit: a factorial code maximizes information transfer,” *Network* **5**, pp. 565–581, 1994.
16. B. Olshausen and D. Field, “Natural image statistics and efficient coding,” *Network: Computation in Neural Systems* **7**(2), pp. 333–340, 1996.