Classifying faces by race and sex using an autoassociative memory trained for recognition

Alice J. O'Toole*, Hervé Abdi*, Ken A. Deffenbacher, James C. Bartlett***

* School of Human Development
The University of Texas at Dallas
Richardson, TX 75083-0688

** Department of Psychology
The University of Nebraska at Omaha
Omaha, NE 68182-0274

Abstract[1]

We examine the ability of an autoassociative memory trained with faces to classify faces by race and by sex. The model learns a low-level visual coding of Japanese and Caucasian male and female faces. Since recall of a face from the autoassociative memory is equivalent to computing a weighted sum of the eigenvectors of the memory matrix, faces can be represented by these weights and the set of corresponding eigenvectors. We show that reasonably accurate classification of the faces by race and sex can be achieved using only these weights. Hence, race and sex information can be extracted in the model without explicitly learning the classification itself.

Introduction

Adaptive interaction with people in daily life requires proficiency in processing faces in a variety of ways. In addition to our ability to recognize faces, we are also proficient at categorizing faces along a number of dimensions including sex, age, and race. Despite the ease with which we perform these tasks, face processing presents a number of difficult computational problems. As spatial patterns, faces are highly similar in that they all contain the same "features", (e.g., nose, mouth, eyes) arranged in roughly the same configuration. Thus, the information available for distinguishing previously seen faces from new faces, and for classifying faces along other important dimensions, must be found in subtle variations in this prescribed feature set and configuration.

Quantification and representation of the information in faces in a way that makes the tasks of recognition and classification possible is a paradoxical problem. The purpose of recognition is to classify a face as familiar or unfamiliar. Thus, a successful computational model of face recognition should be capable of storing a face in a way that makes it distinguishable from faces which were not learned. This needs to be done regardless of the similarity relationships among familiar and unfamiliar faces. It has been known since Kohonen (1977) that an autoassociative memory, composed of faces represented simply by pixel values, acts as a content addressable memory. O'Toole, Millward, & Anderson (1988) used this type of model in a two-alternative forced choice face recognition task and showed that the model distinguished learned from new faces with a good degree of accuracy.

The problem of classification of faces, on the other hand, requires that faces be categorized according to a set of defining characteristics. For simplicity, we will confine the discussion to categorization by race. But, the present approach holds analogously for the problem of face categorization by sex and age as well. In Bruce & Young's (1986) model of face processing, the information needed to do this kind of classification is called "visually-derived semantic". The term makes the point that these categorizations do not require familiarity with the face, but rather are based on perceptual information and so can be done for unfamiliar as well as familiar faces.

To say that certain categorizations of faces are based on visual information, however, does not mean that learning is not important. While it may not be evident in categorizing a face as Caucasian, Black, or Oriental, it is clear that more subtle race categorizations, such as categorizing an oriental face as Chinese, Japanese, or Vietnamese, does require some learning. We believe that perceptual learning is responsible for fine-tuning our ability to discriminate faces both within and between categories (O'Toole et al., in press).

There is abundant support for an other-race effect in which people are better able to recognize faces of their own race than faces of another race (e.g., the meta-analysis of Shapiro & Penod, 1986). O'Toole et al. (in press) modeled the "other-race effect" in face recognition as a problem in perceptual learning. With greater exposure to faces of a "majority" race, the model was able to make finer discriminations between faces within the majority race and was better able to recognize faces (i.e., discriminate learned from new faces) in the majority race than in the minority race.

Returning to the problem of face classification by race, one approach would be to train a connectionist model to do the classification (cf. Cottrell & Fleming, 1990; Golomb, Lawrence, & Sejnowski, 1991), reinforcing the model for correct categorization. The difficulty with this strategy, in its simplest form, is that learning the classification amplifies the differences between face groups at the expense of de-emphasizing the differences between faces

1. Thanks are due to June Chance and Al Goldstein for providing the Caucasian and Japanese faces used in the simulations, and to Peter Assmann and Barbara Edwards for helpful comments on an earlier version of this manuscript.

within a category. This makes the representations of faces within the model less suitable for recognition. Using back-propagation, Cottrell & Fleming (1990) and Golomb et al. (1991) avoid this problem by preprocessing faces with an image compression network before teaching the categorization. Both models perform well on the task of sex discrimination. In the present work, we show that ability to classify faces into one of two race categories is present implicitly in a model that has not been explicitly taught to do this classification. We use an autoassociative memory that is trained for face recognition and show that information about the race and sex of the face can be extracted with a reasonable degree of accuracy.

The Autoassociative Memory. The model is defined first and then its application to the categorization problem is presented. An autoassociative memory was created as follows. A digitized image of each face was coded as a vector of pixel elements concatenated from the rows of the face image. Thus, the ith face was represented by a $J \times 1$ vector (where J equals the number of pixels in the image) denoted by $\mathbf{f}_i$. The vectors were normalized. The faces were stored in an autoassociative memory composed of J completely inter-connected units. The connection strengths were stored in a $J \times J$ matrix $\mathbf{A}$ constructed as follows:

$$\mathbf{A} = \Sigma\, \mathbf{f}_i \mathbf{f}_i^T \tag{1}$$

Recall of individual faces from the matrix was done as:

$$\hat{\mathbf{f}}_i = \mathbf{A}\, \mathbf{f}_i \tag{2}$$

where $\hat{\mathbf{f}}_i$ is the system estimate of $\mathbf{f}_i$. The quality of this estimate is measured by comparing the reconstructed image with the original image. This is done by taking the cosine of the angle between $\hat{\mathbf{f}}_i$ and $\mathbf{f}_i$. Interestingly, the comparison may be done visually by displaying the system output, since it will also be a vector of pixels. System performance was improved by using the Widrow-Hoff error-correction rule that iteratively changes the weights in $\mathbf{A}$ to optimize the quality of the recall across the stimulus set:

$$\mathbf{A}_{(t+1)} = \mathbf{A}_{(t)} - \gamma(\mathbf{f}_i - \mathbf{A}_{(t)}\mathbf{f}_i)\mathbf{f}_i^T \tag{3}$$

where i is randomly chosen, and where γ decreases as an inverse monotonic function of the iteration number t.

While the formulation above shows intuitively what the model is doing, clearly the eigen-decomposition of this matrix is equivalent to principal component analysis (Anderson, et al., 1977; Abdi, 1988). Further, since $\mathbf{A}$ can be expressed as a weighted sum of the outer products of its eigenvectors, recall of a given face from the matrix (i.e, Equation 3) is equivalent to producing a weighted sum of eigenvectors:

$$\begin{aligned}\hat{\mathbf{f}}_i &= \mathbf{A}\mathbf{f}_i = \mathbf{P}\Lambda\mathbf{P}^T\mathbf{f}_i \\ &= \sum_k^N \lambda_k \mathbf{p}_k \mathbf{p}_k^T \mathbf{f}_i = \sum_k^N \lambda_k \cos(\mathbf{f}_i, \mathbf{p}_k)\mathbf{p}_k\end{aligned} \tag{4}$$

where $\mathbf{P}$ is the matrix of the N eigenvectors of $\mathbf{A}$ with eigenvalues greater than zero, and Λ is the diagonal matrix of the eigenvalues in decreasing order, and $\mathbf{p}_k$ is the kth eigenvector. The Widrow-Hoff rule can be expressed via the eigen-decomposition of $\mathbf{A}$ (Abdi, in press):

$$\mathbf{A}_{(t)} = \mathbf{P}[\mathbf{I} - (\mathbf{I} - \gamma\Lambda)^t]\mathbf{P}^T \tag{5}$$

When converged, Equation 4 reduces to (Kohonen, 1977):

$$\begin{aligned}\hat{\mathbf{f}}_i &= \tilde{\mathbf{A}}\mathbf{f}_i = \mathbf{P}\mathbf{P}^T\mathbf{f}_i \\ &= \sum_k^N \mathbf{p}_k\mathbf{p}_k^T\mathbf{f}_i = \sum_k^N \cos(\mathbf{f}_i, \mathbf{p}_k)\mathbf{p}_k\end{aligned} \tag{6}$$

(i.e., the eigenvalues are then equal to one). As the dimensionality of our images is extremely large by comparison to the number of stimuli, the capacity limit of the system was not a problem for these simulations.

Using this latter formulation of the problem, O'Toole & Abdi (1989) have pointed out the similarity of this approach to traditional multidimensional scaling approaches to representing human similarity data. The axes of the multidimensional scaling solution are analogous to the eigenvectors of the autoassociative matrix. The goal of multidimensional scaling is to represent the similarity relations among stimuli using the smallest number of dimensions possible. Frequently, human similarity data can be represented with a good degree of accuracy using only a few dimensions. Analogously, with an autoassociative matrix composed of a pixel-based representation of faces, Sirovich & Kirby (1987) have shown that faces can be reconstructed to a quite recognizable form (to within 3% error) using only 40 parameters and the corresponding 40 eigenvectors. O'Toole et al. (in press) showed that for the task of simply discriminating learned and unlearned faces, even less parameters are needed. Here reconstructions based on as few as ten eigenvectors were sufficient to produce excellent discrimination of learned and unlearned faces. Recently, Turk & Pentland (1991) have elaborated on this approach in several ways including the addition of an algorithm to locate faces in an image.

The purpose of the present work is to examine the usefulness of these coefficients (i.e., the set of weights used for reconstructing a given face) for predicting class membership. Specifically, we will look at classification by race and sex.

Method

Stimuli. One-hundred and sixty-seven faces were digitized from slides to 16 gray levels using a Fotovix digitizer attached to a 286-based computer with a 16-bit TARGA board (True Vision). Faces were of young adults, with roughly half Caucasian faces and half Japanese faces. Within each race set, faces were roughly half male faces and half female faces. None of the slides pictured people

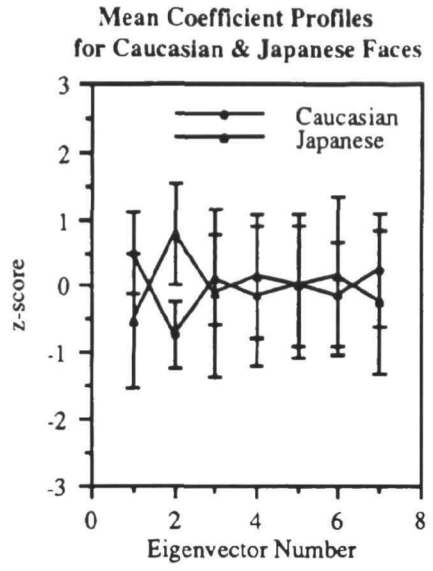

Figure 1. Mean coefficient *z*-score profiles of Caucasian and Japanese faces. Error bars show the standard deviations of the *z*-scores. The best race separation is achieved with the second eigenvector.

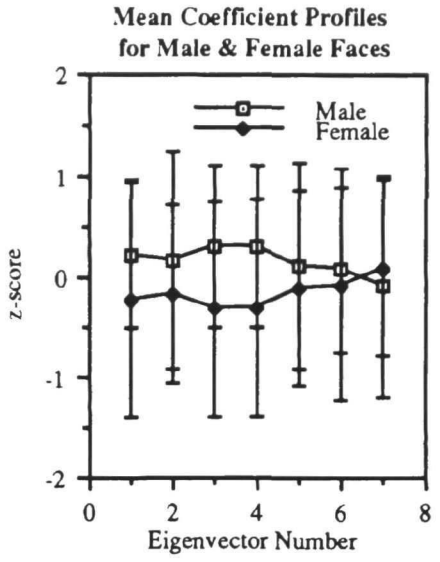

Figure 2. Mean coefficient *z*-score profiles of male and female faces. The error bars indicate the standard deviations of the male and female *z*-scores for each race of faces.

with facial hair or glasses. The images were aligned, so that the eyes of all faces were at about the same height. The images were cropped to eliminate clothing. Each face was 151 pixels wide and 225 pixels long (i.e., a 33,975-pixel vector consisting of the concatenation of the pixels row). We used a spatial differentiation encoding that enhanced lines prior to the extraction of the pixel vector, (cf. O'Toole et al., 1988).

Apparatus. Simulations were performed on a Sun SparcStation and a Convex C-1 Vector computer.

Procedure. An autoassociative memory was created using 40 Caucasian and 40 Japanese faces. The eigenvectors were extracted from the matrix and reconstructions of all 167 faces (i.e., 80 old faces and 87 new faces) were made using Equation 4, with the first seven[2] eigenvectors.

Analysis and Discussion - Race Discrimination. A coefficient profile was constructed for each face using the seven coefficients used to reconstruct the face. Since the absolute magnitudes of these coefficients were different for different eigenvectors, the face profiles were converted into *z*-scores. Thus, each coefficient in a face profile was the *z*-score of the coefficient with respect to that coefficient for all faces. An average coefficient profile was then calculated for Caucasian faces and for Japanese faces. The mean coefficients for Japanese and Caucasian faces were significantly different for the first [t (165) = 7.87, $p < .001$], second [t(165) = 15.25, $p < .001$], and seventh [t (165)= 3.32, $p < .01$] eigenvectors. This indicates that information relevant to race discrimination is present in each of these eigenvectors. Figure 1 displays these average profiles along with error bars indicating one standard deviation.

We looked at the power of the coefficients for eigenvectors one and two, independently, for making race predictions about the faces. The simplest scheme for doing this is to take the mean of the mean coefficients for the Japanese and Caucasian faces, and to create a decision rule using the mean of these two means as a criterion. Race membership predictions are made by assigning faces with coefficients exceeding this criterion to one race (i.e., the race with the larger of the two coefficient means) and faces with coefficients less than this criterion to the other race (i.e., the race with the smaller of the two coefficient means). Using only the coefficients for eigenvector one yielded correct race predictions for 74.8% of the faces. Basing decisions on the coefficients for the second eigenvector fared better, yielding correct race predictions for 88.6% of the faces.

Analysis and Discussion - Sex Discrimination. Average face profiles were created for male and female faces using the procedure described above for race profiles. Figure 2 shows this analysis. These data show that the information for sex discrimination in the model is not localized in one or two eigenvectors. The mean coefficients for male and female faces were significantly different for the first [t (165) = 3.01, $p < .01$], second [t (165) = 2.29, $p < .05$], third [t(165) = 4.21, $p < .001$], and fourth eigenvectors [t (165) = 4.07, $p < .001$]. Figure 2 indicates that with the large overlap of the mean distributions, no eigenvector

2. The choice of seven was made somewhat arbitrarily to yield moderate recognition performance. The model discriminated learned from new faces with a d' of 1.08 using seven eigenvectors.

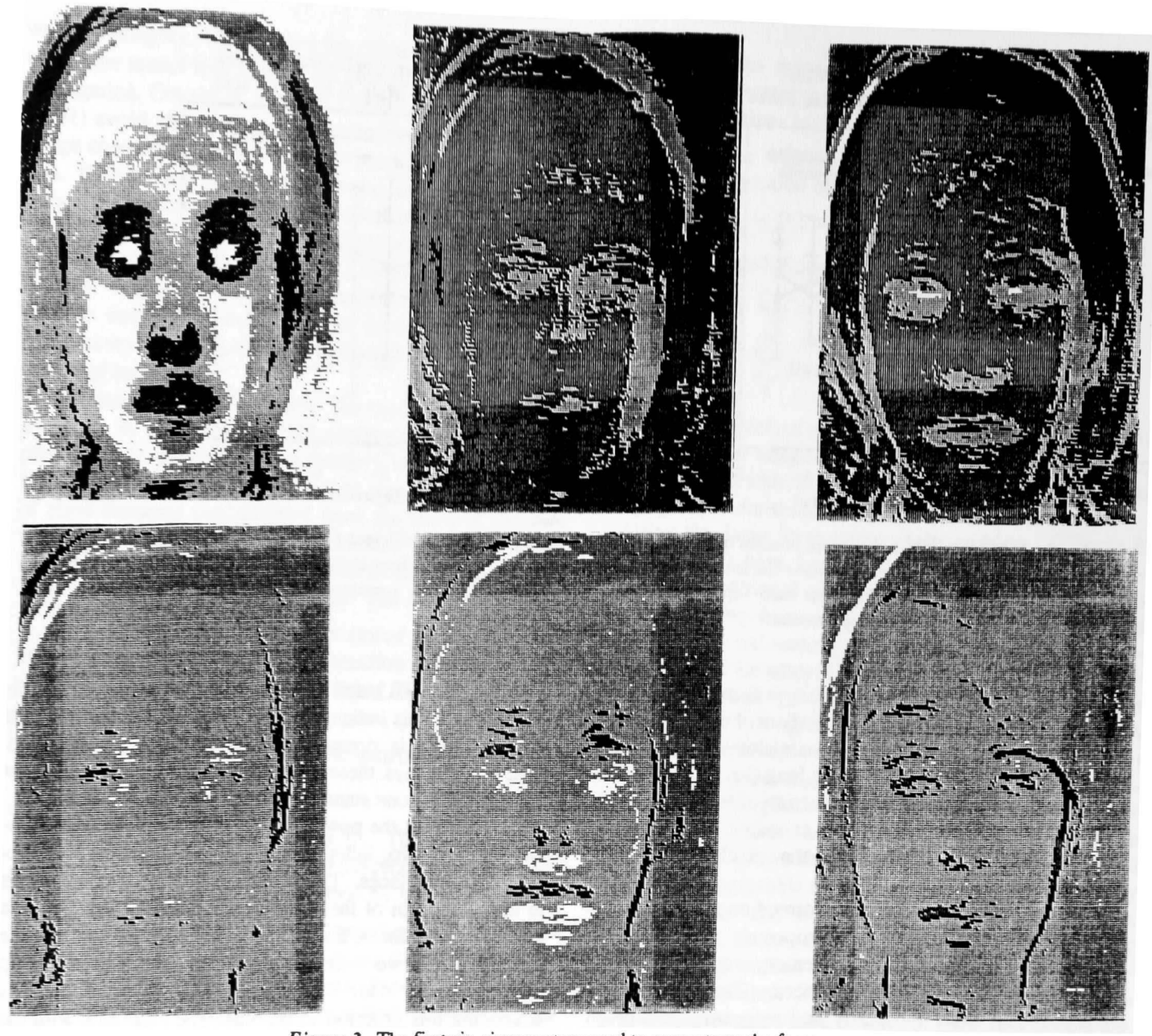

Figure 3. The first six eigenvectors used to reconstruct the faces.

independently will be sufficient to discriminate male and female faces. Since for all of these eigenvectors the mean coefficient for male faces was greater than for female faces, the simplest method for combining information across eigenvectors was to add the coefficients for the four eigenvectors for each face and to use this to predict sex classification. Again, we took the mean of the means of this predictor for male and for female faces and used it as a criterion sex value, assigning male to faces exceeding the criterion value and female to faces less than the criterion value. Using the sum of these coefficients yielded 74.3% correct sex predictions.

Eigenvectors as Features. Since the face codings are pixels vectors, the eigenvectors of the memory matrix are also pixel vectors and some examples appear in Figure 3. As noted previously (Abdi, 1988; O'Toole & Abdi, 1989), these vectors are face-like. Further, since any face can be expressed as a weighted combination of eigenvectors, the eigenvectors qualify in some ways as "features" of the face with the coefficients indicating the degree of presence of the feature. The idea of using eigenvectors as features has been around for a relatively long time (Anderson, et al., 1977). These are not localized features but rather more global features. Since the eigenvectors reflect the statistical structure of the stimulus set, a more race-homogeneous matrix will yield eigenvectors that are different from those we have extracted from a matrix made of equal numbers of stimuli from two races. For example, Figure 4 shows the first eigenvector from an autoassociative matrix created from 95% Caucasian and 5% Japanese faces and from 95%

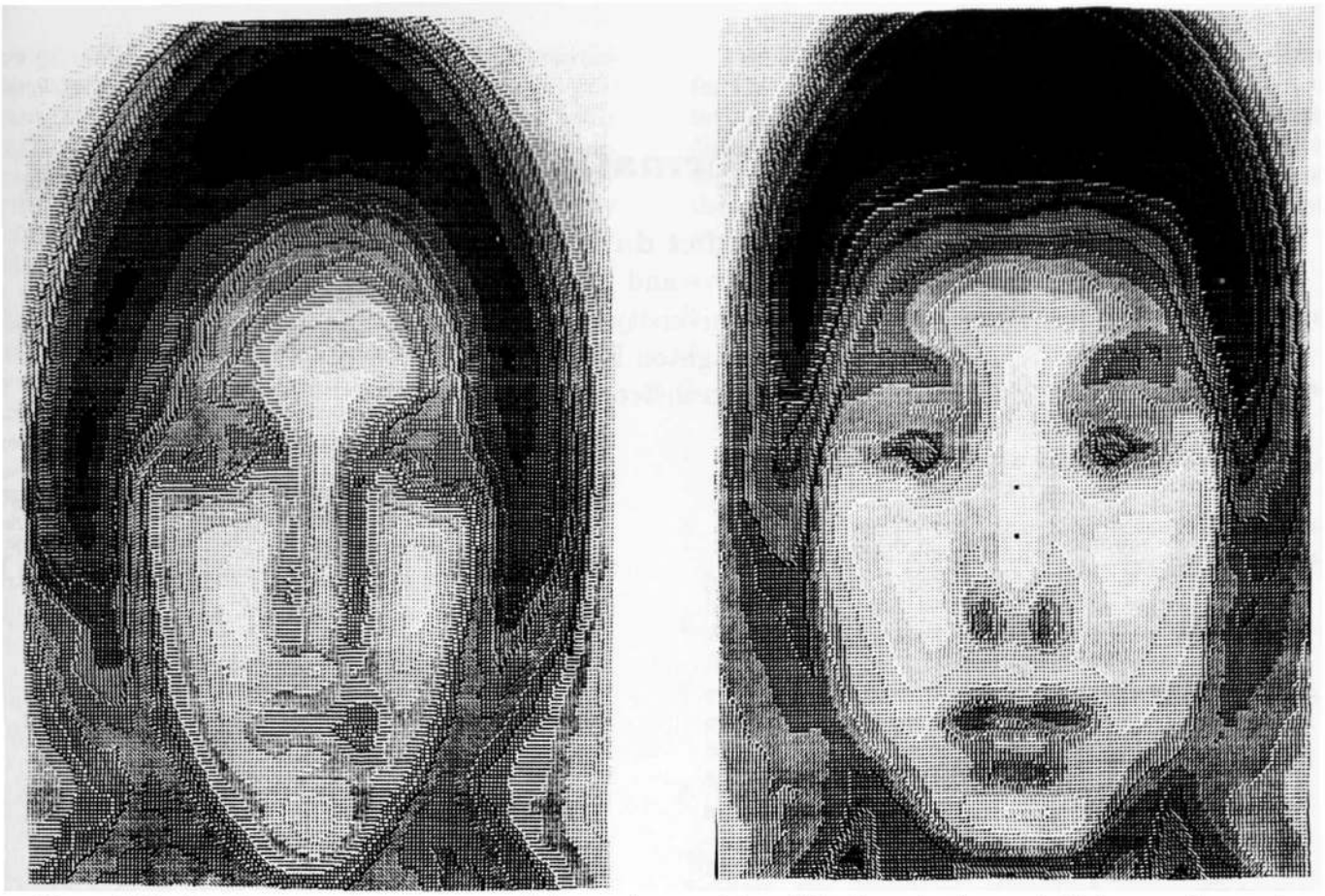

Figure 4. The first eigenvector from an autoassociative matrix created from 95% Caucasian faces and 5% Japanese faces and from 95% Japanese faces and 5% Caucasian faces , respectively.

Japanese and 5% Caucasian faces, respectively. Here the first eigenvectors are typical of the "majority" race in the matrix.

Summary. This work shows that the information needed to make race and sex classifications is present in a model trained to perform face recognition. While the discrimination ability of the model is not perfect, it is well above chance. We think that the competing concerns of recognition and classification should be addressed with a unified representation of the stimuli. Future work needs to address the question of how a single model can represent stimuli in a way that answers competing task needs.

References

Abdi, H. 1988. A generalized approach for connectionist auto-associative memories: interpretation, implications and illustration for face processing. In J. Demongeot (Ed.), *Artificial Intelligence and Cognitive Sciences*. Manchester: Manchester University Press.

Abdi, H. in press. *Les Réseaux de Neurones*. Presses Universitaires de Grenoble .

Anderson, J. A., Silverstein, J. W., Ritz, S. A., Jones, R. S. 1977. Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review, 84*, 413-451.

Bruce,V., Young, A. W. 1986. Understanding face recognition. *British Journal of Psychology, 77*, 305-327.

Cottrell, G. W., Fleming, M. 1990. Face recognition using unsupervised feature extraction. *IJCNN-90, 2*, 65-70

Golomb, B. A., Lawrence, D. T., Sejnowski, T. J. 1991. SEXnet: A neural network identifies sex from human faces. In D. S. Touretsky, R. Lippmann (Eds.) *Advances in Neural Information Processing Systems, 3*, CA:San Mateo: Morgan Kauffmann.

Kohonen, T. 1977. *Associative memory: A System Theoretic Approach*. Berlin: Springer-Verlag.

O'Toole, A. J., Abdi, H. 1989. Connectionist approaches to visually based feature extraction. In G. Tiberghien (Ed.) *Advances in Cognitive Psychology*, (Vol 2). London: John Wiley.

O'Toole, A. J., Deffenbacher, K. A., Abdi, H., Bartlett, J. A. in press. Simulating the "other-race effect" as a problem in perceptual learning. *Connection Science*.

O'Toole, A. J., Millward, R. B., Anderson, J. A. 1988. A physical system approach to recognition memory for spatially transformed faces. *Neural Networks, 1*, 179-199.

Shapiro, P.N., Penod, S.D. 1986. Meta-analysis of face identification studies. *Psychological Bulletin, 100*, 139-56.

Sirovitch, L., Kirby, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America, 3*, 519-524.

Turk, M., Pentland, A. 1991. Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*, 71-86.