[10] O'Toole A.J. and Abdi H., Connectionist approaches to visually based feature extraction. In *Advances in Cognitive Psychology* **2**, ed. by Tiberghien G. (John Wiley, London, 1989) pp. 124–140.

[11] O'Toole A.J., Abdi H., Deffenbacher K.A. and Bartlett J.C., Classifying faces by race and sex using an autoassociative memory trained for recognition. In *Proc. 13th Annu. Conf. Cognitive Sci. Soc.*, ed. by Hammond K.J. and Gentner D. (Erlbaum, Hillsdale, 1991) pp. 847–851.

[12] O'Toole A.J., Abdi H., Deffenbacher K.A. and Valentin D., Low-dimensional representation of faces in higher dimensions of the face space. *J. Opt. Soc. Am. A* **10** (1993) 405–410.

[13] O'Toole A.J., Deffenbacher K.A., Abdi H. and Bartlett J.C., Simulating the other race effect as a problem in perceptual learning. *Connection Sci.* **3** (1991) 163–178.

[14] O'Toole A.J., Millward R.B. and Anderson J.A., A physical system approach to recognition memory for spatially transformed faces. *Neural Networks* **1** (1988) 179–199.

[15] Sirovich L. and Kirby M., Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A* **4** (1987) 519–524.

[16] Solheim I, Payne T. L. and Castain R. C. The potential in using back-propagation neural networks for facial verification system. WINN - AIND, Auburn, Al (1992).

[17] Turk M. and Pentland A., Eigenfaces for recognition. *J. Cognitive Neurosci.* **3** (1991) 71–86.

[18] Valentin D., Abdi H., O'Toole A.J., & Cottrell G.W. Connectionist models of face processing: A survey. *Pattern Recognition* **27** (in press).

[19] Valentin D., Abdi H. and O'Toole A.J., Principal component and neural network analyses of face images: Explorations into the nature of information available for classifying faces by sex. (submitted).

5. Concluding remarks

We have shown that complex perceptual tasks such as face identification and categorization can be successfully performed by simple linear models used in conjunction with a pixel-based coding of the faces. These models analyze faces in terms of the eigendecomposition of a pixel cross-product matrix, as such, the representation (*i.e.,* "macrofeatures"or eigenvectors) they use to describe the faces reflects the statistical structure of the set of faces on which they are trained. Although not intended as a solution to the general problem of face processing, linear models constitute an interesting tool for analyzing the properties of faces as complex visual patterns, and hence, can lead to new insights into the way human beings are able to process faces so effortlessly and efficiently. Further work needs to concentrate on the psychological relevance of those insights.

6. Acknowledgment

Thanks are due to June Chance and Al Goldstein for providing the faces used in the simulations and to Betty Edelman for helpful comments on earlier drafts of this paper.

References

[1] Abdi H., A generalized approach for connectionist auto-associative memories: interpretation, implications and illustration for face processing. In *Artificial Intelligence and Cognitive Sciences*, ed. by Demongeot J. (Manchester University Press, Manchester, 1988) pp. 151–164.

[2] Abdi H., *Les Réseaux de Neurones* (Presses Universitaires de Grenoble, Grenoble, 1994).

[3] Abdi H., Valentin D. and O'Toole A.J., More about the difference between men and women: Evidence from linear neural networks and principal component approaches. (in preparation).

[4] Anderson J.A., Silverstein J.W., Ritz S.A. and Jones, R.S., Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychol. Rev.* **84** (1977) 413–451.

[5] Burton M. and Bruce V., What's the difference between men and women? Evidence from facial measurement. *Perception* **22** (1993) 153–176.

[6] Cottrell G.W. and Fleming M.K., Face recognition using unsupervised feature extraction. In *Proc. Int. Conf. Neural Network* (Kluwer, Dordrecht, 1990) pp. 322–325.

[7] Golomb B.A., Lawrence D.T. and Sejnowski T.J., Sexnet: a neural network identifies sex from human faces. In *Advances in Neural Information Processing System* **3**, ed. by Lippman R.P, Moody J. and Touretzky D.S. (Morgan Kaufman, San Mateo, 1991) pp. 572–577.

[8] Kohonen T., *Associative memory: A system theoretic approach.* (Springer-Verlag, Berlin, 1977).

[9] Millward R. and O'Toole A., Recognition memory transfer between spatial-frequency analyzed faces. In *Aspects of Face Processing*, ed. by Ellis H.D., Jeeves M.A., Newcombe F. and Young A. (Martinus Nijhoff, Dordrecht, 1986) pp. 34–44.

When an eigendecomposition pre-processing of the faces was used, an autoassociative memory was first created to store the faces of the training set and then decomposed into its eigenvectors and eigenvalues. The projections of both learned and new faces onto the eigenvectors were then computed. The projections of the learned faces were used to trained either a perceptron or a radial basis function network to produce a 0 for male faces and a 1 for female faces. After learning was completed, the projections of the new faces were used to test the ability of the classification networks to generalize sex classification to new faces. Note that the new faces were not used to compute the eigenvectors nor the optimum weights of the classification networks. When no pre-processing was used, the classification networks were trained to classify the faces from the training set using the pixel intensity vectors directly as input. The ability of the classification networks to generalize was then tested by presenting new faces as input to the networks.

To maximize the number of new faces, the performance of the networks was evaluated using a "leave-one-out" or jackknife technique. In brief, having a data set of 160 faces, training was performed on 159 faces leaving one face out for testing. After learning was completed, the network was used to predict the sex of the test face. Then the input set was alternated and the procedure repeated so that each of the 160 faces was used, in turn, as the "new face". Results showed: 1) no difference between the two classification networks: a simple perceptron performed as well as the more complex radial basis function network, indicating that sex categorization is mostly a linear problem; 2) a better level of performance was found for both old and new faces when the face images were pre-processed *via* an eigendecomposition (100% *vs* 81% for old faces and 91% *vs* 80% for new faces). This latter result indicates that the eigendecomposition step is important in achieving a level of performance somewhat comparable to human performance in a sex categorization task.

In summary, the Abdi *et al.* results suggest that pre-processing the faces using a linear autoassociative memory not only saves processing time by reducing the size of the sex classification network (from 33975 to 160 input units), but also produces a set of features relevant for discriminating between female and male faces. This has the major advantage of eliminating the difficult problem encountered by Burton *et al.* [5] of finding, *a priori*, a set of features useful for categorizing faces according to their sex.

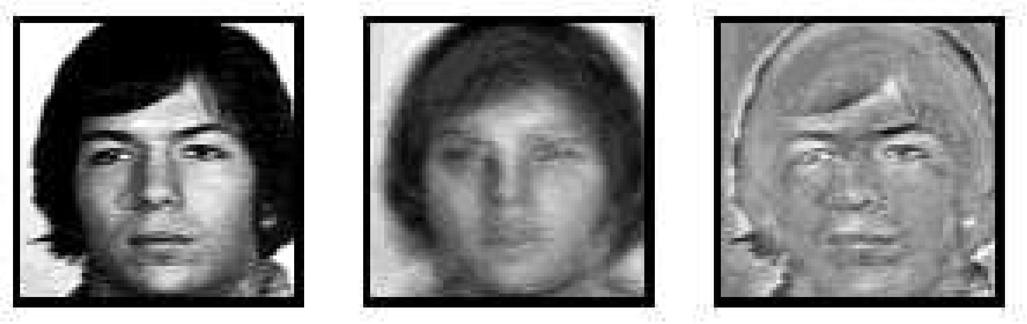

FIG. 10. — Illustration of the different kinds of facial information conveyed by different ranges of eigenvectors. From left to right: the original face and reconstruction using 1) the first 30 eigenvectors [$r^2 = .90$], 2) all but the first 30 eigenvectors [$r^2 = .10$].

proportion of the variance in the face set, and because they code what is unique to individual faces, these eigenvectors are crucial for distinguishing a particular face from any other face. This "dissociation" between categorical and identity specific information is illustrated by Figure 10. The left panel of this figure displays the original face, the middle panel a reconstruction of the face with the first 30 eigenvectors, and the right panel a reconstruction of the face eliminating the first 30 eigenvectors. While the reconstruction with the first 30 eigenvectors explained most of the variance in the face ($r^2 = .90$), most human observers would have a problem in identifying it, but would be able to indicate the sex and race. In contrast, the second reconstruction (eigenvectors 31 to 160) explained only 10% of the total variance but provides enough information for identifying the face.

4. Categorizing faces according to their sex

In addition to identifying and recognizing familiar faces, human beings are able to categorize unfamiliar faces along visually derived dimensions (*e.g.,* sex, race, or age). Among these perceptual categorizations, sex classification is one of the most biologically important and probably the easiest to achieve. For example, Burton, Bruce, and Dench [5], reported that human subjects were able to classify photographs of 179 adults with respect to sex with 96% accuracy even though the hair was concealed by a swimming cap. Yet, despite the apparent ease with which human beings are able to separate male from female faces, it is not easy to find a set of simple measurement-based discriminators allowing a level of sex discrimination equivalent to the one obtained by human subjects [5].

Recently, Abdi, Valentin, and O'Toole [3] examined the usefulness of the "macrofeatures" extracted by an autoassociative memory, trained to reconstruct a set of faces, in simulating a sex categorization task. They trained two classification networks (a standard perceptron and a radial basis function network) to classify a set of face images according to their sex. The face images were either pre-processed using a linear autoassociative memory or used directly as input to the classification networks.

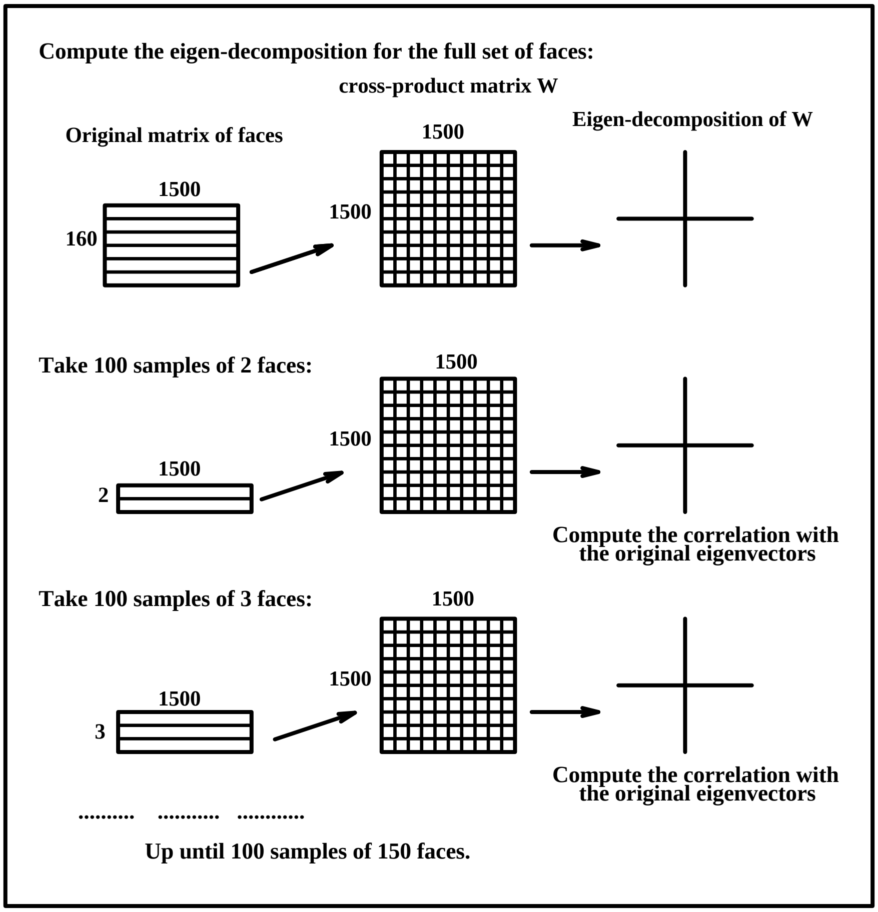
Compute the eigen-decomposition for the full set of faces:
cross-product matrix W
Original matrix of faces
1500
160
1500
1500
Eigen-decomposition of W
Take 100 samples of 2 faces:
1500
2
1500
1500
Compute the correlation with the original eigenvectors
Take 100 samples of 3 faces:
1500
3
1500
1500
Compute the correlation with the original eigenvectors
....
Up until 100 samples of 150 faces.

FIG. 9. — Algorithm used by Valentin et al. [19] to estimate the stability of the eigenvectors extracted from a cross-product matrix of face images.

used (*i.e.,* when the eigenvalues are not equalized) the response of the memory to almost any memory key is the first eigenvector (cf. Figure 2).

3.4. Dissociation between categorical and identity specific information

In summary, the studies presented above suggest that different kinds of facial information are conveyed by different ranges of eigenvectors. These different types of information have different properties and are not equally useful for all tasks. The eigenvectors with large eigenvalues seem to convey mainly low frequency information (*e.g.,* basic shape and structure of the face), they are very stable, and can be easily computed [2]. When heterogeneous sets of faces (*e.g.,* faces of different sexes and races) are used, these eigenvectors contain reliable information for making visually-based categorizations. In contrast, the eigenvectors with small eigenvalues seem to convey essentially spatial high-frequency information (*e.g.,* specific shape of the eyes, nose, and mouth), and hence, are very unstable. Despite the fact that they explain only a very small

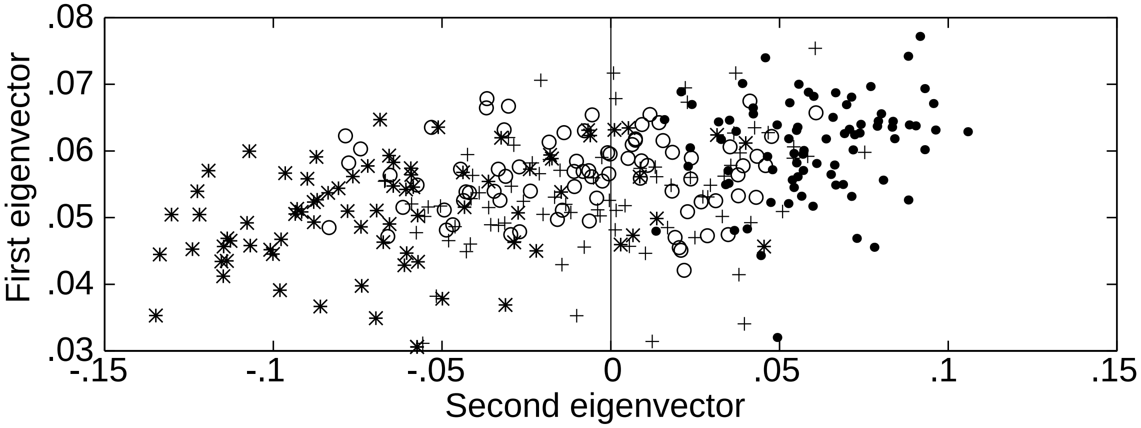

FIG. 8. — Projections of face images onto the space determined by the first two eigenvectors of an autoassociative memory created from a set of 320 face images: 80 male Caucasian (○), 80 female Caucasian (*), 80 male Japanese (●), and 80 female Japanese (+). Clearly, the second eigenvector opposes the male Japanese faces to the female Caucasian faces but does not separate female Japanese faces from male Caucasian faces. Note that if you look at the race dimension independently of sex, the second eigenvector on the whole allows for a good discrimination between Caucasian and Japanese faces.

3.3. Eigenvector stability

A recent study by Valentin, Abdi, and O'Toole [19], evaluated the stability of the information carried by different eigenvectors as a function of their eigenvalues. They were interested in finding the minimum number of faces necessary to estimate accurately the eigenvectors of an autoassociative memory created from 160 Caucasian face images. The algorithm used to estimate the stability of the eigenvectors is illustrated in Figure 9. First, an autoassociative memory was created using the complete set of face images (160 faces: 80 males and 80 females). This memory was decomposed into its eigenvectors. Then, face samples ranging in size from 2 to 150 (100 samples for each size condition) were randomly selected from the original set of faces. For each sample, an autoassociative memory was created and decomposed into its eigenvectors. Finally, a coefficient of correlation was computed between the eigenvectors extracted from the face samples and the original eigenvectors.

Results showed that the minimum number of faces necessary to estimate correctly the original eigenvectors varies as an inverse function of their eigenvalues. Specifically, the pattern of stability reported by Valentin *et al.* indicated a very high stability of the first eigenvector ($r^2 = .88$ with as few as 2 faces), a lesser but still good stability of the next five eigenvectors ($r^2 = .80$ with as few as 50 faces) and a decreasing stability of the eigenvectors with smaller eigenvalues. The extreme stability of the first eigenvector is not really surprising since this eigenvector captures the information that is most common to all faces, and hence, remains stable across samples. The eigenvalue associated with this eigenvector is very large (81% of the total inertia). This extreme stability explains the fact that when standard Hebbian learning is

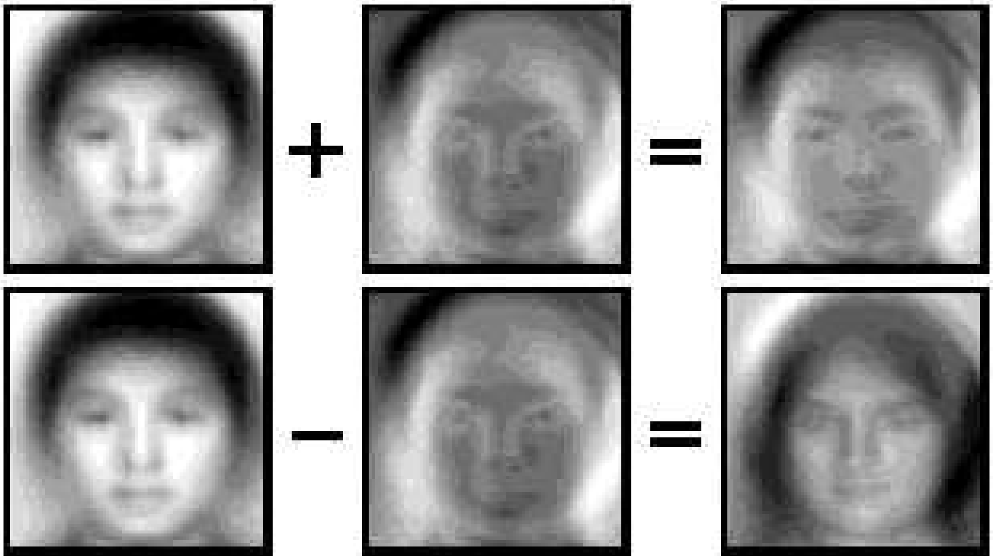

FIG. 7. — Illustration of the role of the second eigenvector in determining the racial and sexual appearance of a heterogeneous set of faces. *Top panels*: Adding the first eigenvector to the second eigenvector creates a male Japanese looking face. *Bottom panels*: Subtracting the second eigenvector from the first eigenvector creates a female Caucasian looking face.

The importance of the second eigenvector in determining the racial and sexual appearance of a heterogeneous set of faces is illustrated in Figure 7. This illustration was inspired by an earlier demonstration [12] of the role of the second eigenvector in determining the sex of faces of a relatively homogeneous set of male and female faces (*i.e.*, composed of a single race of faces). In this previous demonstration, O'Toole *et al.* showed that adding the first eigenvector to the second eigenvector creates a masculine looking face whereas subtracting it from the second eigenvector creates a feminine looking face. In the demonstration presented here the memory was trained to reconstruct a set of 320 face images comprised of 160 Caucasian faces (80 males and 80 females) and 160 Japanese faces (80 males and 80 females) and decomposed into its eigenvectors. Figure 7 shows that adding the second eigenvector to the first eigenvector creates a male Japanese looking face, in contrast, subtracting it from the first eigenvector creates a female Caucasian looking face. From a purely statistical point of view, this can be explained by the fact that the second eigenvector opposes the male Japanese faces to the female Caucasian faces. In other words, male Japanese and female Caucasian faces load strongly and in opposite directions on the second eigenvector whereas female Japanese and male Caucasian have similar loadings in the middle part of this eigenvector (cf. Figure 8).

3.2. Perceptual information in faces

An additional advantage of the eigenvector approach to linear autoassociators is that it provides a tool for analyzing the perceptual information in faces. Recently, O'Toole, Abdi, Deffenbacher and Valentin [12] examined the kind of information that is provided by different ranges of eigenvectors and the usefulness of this information with respect to psychological tasks. They showed that different tasks make different demands in terms of the information that needs to be processed, and that this information is not contained in the same ranges of eigenvectors.

In a first simulation, they examined the importance of different ranges of eigenvectors for discriminating learned faces from unlearned ones (recognition task). An autoassociative matrix was created from 100 face images (50 females and 50 males) and decomposed into its eigenvectors. Different ranges of 15 eigenvectors (sorted in decreasing order according to their eigenvalues) were used to reconstruct the 100 learned (or "old") faces and 59 new faces. The recognition task was simulated using a methodology similar to the one described previously [13]. The results indicate that: 1) the quality of representation (physical similarity as measured by a cosine taken between the original and the reconstructed face vectors) decreased as the range of eigenvectors used was shifted from the eigenvectors with larger eigenvalues to those with smaller eigenvalues; 2) the ability of the model to discriminate between old and new faces does not follow the decrement in quality of representation. Any of the 15-eigenvector ranges between the 45-th and the 80-th eigenvectors provided better information for discriminating learned from unlearned faces (*i.e.,* "recognizing" faces) than did the first 15-eigenvector range.

In addition, they examined the memory's ability to predict the sex of faces across the eigenvectors. This was done by computing a point biserial correlation between the sex of the faces (coded as 0 and 1) and their projections on individual eigenvectors. The results showed that the eigenvectors with larger eigenvalues were more useful for predicting the sex of the faces than the eigenvectors with smaller eigenvalues. In particular, the second eigenvector was the best predictor of the sex of the face ($r^2 = .38$). The fact that eigenvectors with relatively large eigenvalues contain information relevant to categorical information has also been reported in a previous study by O'Toole, Abdi, Deffenbacher, and Bartlett [11]. In that study, they showed that when a training set composed of Japanese and Caucasian faces was learned by an autoassociative memory, the eigenvectors with large eigenvalues capture most of the information useful to predict the race and sex of the faces. Specifically, O'Toole *et al.* reported that for their particular training set, the second eigenvector, by itself, yielded correct *race* predictions for 88.6% of the faces, and the sum of the first 4 eigenvectors yielded correct *sex* predictions for 74.3% of the faces. An interesting aspect of these results is that, in both studies [11, 12], the linear autoassociator was not trained to categorize the faces but rather was trained to reconstruct them. In other words, the information necessary to classify the faces according to their race or sex emerged spontaneously from the memory representation.

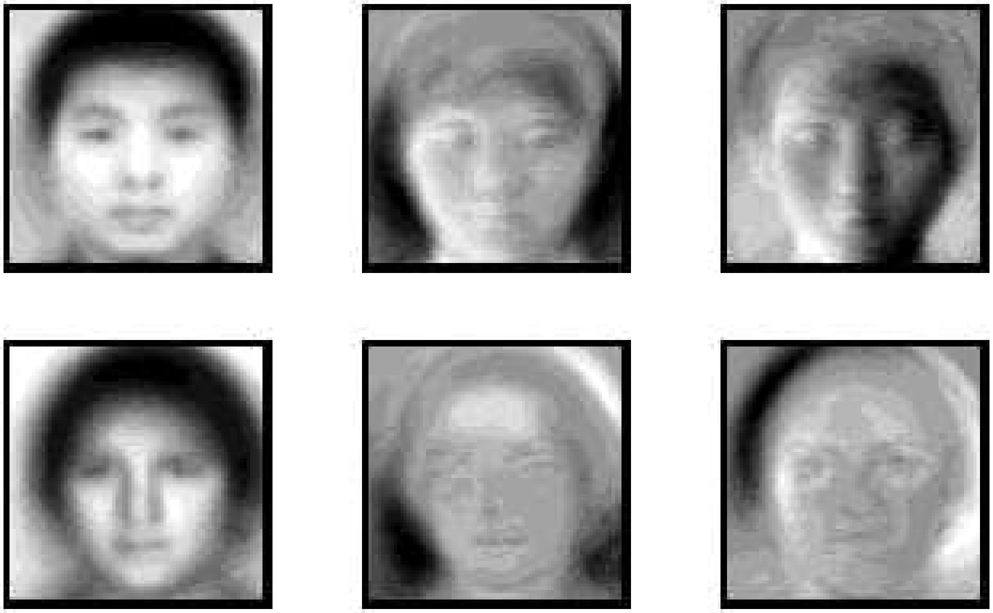

FIG. 6. — *Top panels*: the first 3 eigenvectors extracted from a Japanese face memory. *Bottom panels*: the first 3 eigenvectors extracted from a Caucasian face memory. Global differences in shape and form can be observed between the two sets of eigenvectors.

set of face images. The first memory (referred to as "biased long term memory") was trained to reconstruct a large number of faces of one race ("own race") and a smaller number of faces of another race ("other race"). The second memory (referred to as "unbiased episodic memory") was trained to reconstruct a set of face images composed of half "own race" and half "other race" faces. Old and new faces from majority and minority races were then reconstructed using the eigenvectors extracted from both the biased long term memory and the unbiased episodic memory. The quality of reconstruction of each face was estimated by taking the cosine between the original and the reconstructed face. For the recognition task, the average cosine was taken as a decision criterion. Faces with a cosine above the criterion were categorized as "old" and faces with a cosine below this criterion were categorized as "new". The results showed that faces from the majority race were better recognized than faces from the minority race. In addition, an analysis of the quality of reconstruction of the faces, using only the eigenvectors from the biased autoassociative memory, indicated that: 1) new faces from the majority race were better reconstructed than new faces from the minority race; 2) the memory reconstructions of new faces from the minority race were more similar to each other than the memory reconstructions of new faces from the majority race.

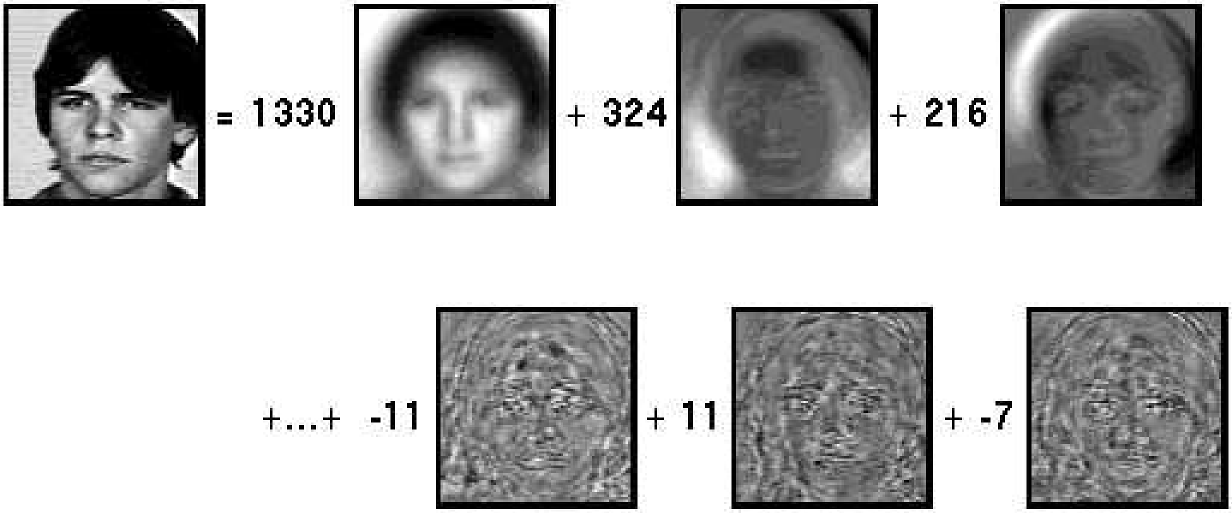

FIG. 5. — Illustration of the reconstruction of a face as a weighted sum of eigenvectors. The weights are the projections of the face on each eigenvector $(\mathbf{u}_\ell^T \mathbf{x}_k)$. They indicate the extent to which a given eigenvector characterizes the face.

where the weights (*i.e.*, $\mathbf{u}_\ell^T \mathbf{x}_k$) are the projections of the faces onto each of the eigenvectors. These weights can be interpreted as an indication of the extent to which a given eigenvector (or "macrofeature") characterizes a particular face (cf. Figure 5).

An advantage of representing faces in terms of eigenvectors is that the eigenvectors reflect the statistical structure of the set of faces from which they are extracted. As an illustration, Figure 6 displays the first 3 eigenvectors extracted from a set of 160 Japanese faces (top panels) and the first 3 eigenvectors extracted from a set of 160 Caucasian faces (bottom panels). It should be noted that these two sets of eigenvectors contain characteristics typical of the race of the face matrix from which they are extracted (*e.g.*, roundness of the face for the Caucasian first eigenvector and squareness of the face for the Japanese first eigenvector).

From a neural network point of view, the system acts somewhat like a Wiener filter [2]. When new faces are presented as memory keys, they are filtered through the features extracted from the set of learned faces. The properties of the filter are determined by the "face history" of the memory matrix. For example, if an autoassociative memory is trained to learn a heterogeneous set of faces, made up of a majority of faces of one race and a minority of faces of another race, it will develop a set of weights optimal for distinguishing between the faces of the majority race (own race) but not for the minority race (other race). In other words, the internal representation (eigenvectors) developed by the model will be determined mostly by the majority race of the training set faces.

O'Toole, Deffenbacher, Abdi, and Bartlett [13] used this property of autoassociative memories to model the well known "other-race effect" as a problem in perceptual learning. To simulate an episodic memory task, they trained two autoassociative memories to reconstruct a

3. Analyzing face images with a linear autoassociator

Both Kohonen [8] and Millward and O'Toole [9] showed that a simple autoassociative memory is a useful tool for modeling face processing. More recent work has attempted additional analyses of the properties of this type of memory. Specifically, these papers try to describe and quantify the perceptual information in faces, and to model the learning of this information. This type of approach, generally referred to as the "principal component approach" to face modeling, relies essentially on the fact that using a linear autoassociator to store and recall face images is equivalent to computing the principal component analysis of the cross-product matrix of a set of faces and reconstructing the faces as a weighted sum of eigenvectors [1].

3.1. *Eigenvectors as macrofeatures*

As pointed out by Anderson, Silverstein, Ritz and Jones [4] and Kohonen [8], and detailed in the paper of Abdi in this volume, since the weight matrix $\mathbf{W}$ is positive semi-definite, it can be expressed as a weighted sum of its eigenvectors:

$$\mathbf{W} = \sum_{\ell=1}^{L} \lambda_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \qquad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{3.1}$$

where $\mathbf{I}$ stands for the identity matrix, $\mathbf{\Lambda}$ represents the diagonal matrix of eigenvalues and L is the rank of the matrix $\mathbf{W}$. The eigenvectors in $\mathbf{U}$ are generally ordered according to their eigenvalues. In what follows, the eigenvector with the largest eigenvalue is referred to as the first eigenvector, the eigenvector with the second largest eigenvalue is referred to as the second eigenvector, and so on.

Similarly, the Widrow-Hoff learning rule can be rewritten in terms of the eigenvectors and eigenvalues of the weight matrix $\mathbf{W}$ at a given time t:

$$\mathbf{W}^{(t)} = \mathbf{U}\mathbf{\Phi}^t\mathbf{U}^T \qquad \text{with} \quad \mathbf{\Phi}^t = [\mathbf{I} - (\mathbf{I} - \eta\mathbf{\Lambda})^t] \ . \tag{3.2}$$

This formulation shows clearly that the Widrow-Hoff error correction rule affects only the eigenvalues of $\mathbf{W}$ [2]. More specifically, it equalizes all the eigenvalues, or in other words, sphericizes the weight matrix. Hence, at convergence, $\mathbf{W}$ reduces to:

$$\mathbf{W}^{\infty} = \mathbf{U}\mathbf{U}^T \ . \tag{3.3}$$

When displayed visually, the eigenvectors of the weight matrix span the entire image and appear face-like. From a psychological point of view, they can be thought of as a set of "global features" or "macrofeatures" from which the faces are built [10, 14, 15]. More formally, a face can be represented as a weighted sum of eigenvectors

$$\hat{\mathbf{x}}_k = \sum_{\ell=1}^{L} \mathbf{u}_\ell \mathbf{u}_\ell^T \mathbf{x}_k = \mathbf{U}\mathbf{U}^T\mathbf{x}_k \ . \tag{3.4}$$

reconstructed quite well but not perfectly ($r^2 = .82$). The new Japanese face is more distorted by the memory than the new Caucasian face ($r^2 = .72$). Finally, the response obtained for the random vector becomes less and less face-like ($r^2 = .00$, after complete Widrow-Hoff learning).

2.2. Model performance

Kohonen [8] was the first to use an autoassociative memory to store and recall face images. Using a sample of 100 faces, he demonstrated that an autoassociative memory could act as a content addressable memory for face images. In his demonstration, an autoassociative memory was first created by autoassociating the face images using a simple Hebbian learning rule (Eq. 2.1). The efficiency of the memory was then tested by presenting noisy or incomplete face images as input. The quality of the reconstructions was estimated by displaying the images reconstructed by the memory. Results showed that the images reconstructed by the system were convincingly similar to the original images. When incomplete or partially obliterated images were presented as memory keys, the memory filled in the missing parts of the images.

Using a similar approach, Millward and O'Toole [9], showed that autoassociative memories can act as an efficient system for recognizing faces (*i.e.,* distinguishing learned from unlearned faces). In their study, an autoassociative memory was constructed by autoassociating a set of face vectors using a Widrow-Hoff learning rule (Eq. 2.4). Face recognition was simulated by using a standard psychological memory paradigm called a "two-alternative forced choice task" (2AFC). This was done by testing the memory with pairs of face vectors, where each pair was composed of an old face (*i.e.,* previously learned by the memory) and a new face (*i.e.,* not learned by the memory). For each face vector (old and new), the quality of response of the model was estimated by computing the cosine between input and output vectors (Eq. 2.3). The face in the pair with the highest cosine was said to be "recognized" by the memory. The results of the 2AFC showed that the linear autoassociator was able to "recognize" faces. On the average, the quality of reconstruction of the faces was higher for the learned faces than for the new faces. Moreover, the memory was able to mimic the pattern of recognition transfer errors found with human subjects performing a recognition transfer task for spatially filtered faces.

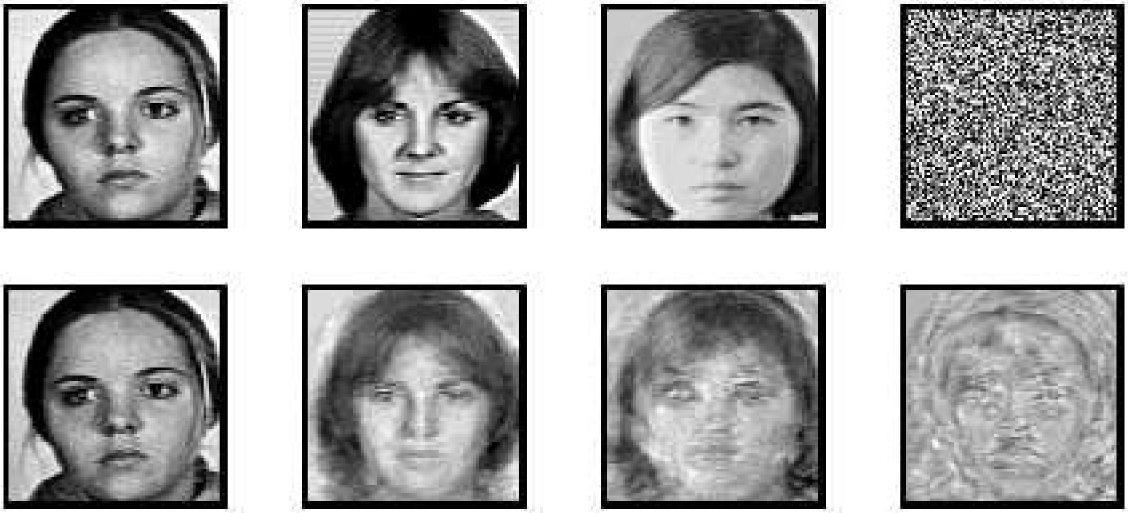

FIG. 4. — The top panels show four stimuli presented as memory keys and the bottom panels the responses produced by an autoassociative memory trained with 160 Caucasian faces. From left to right the stimuli are 1) a Caucasian face that has been learned by the memory, 2) a Caucasian face that has not been learned by the memory, 3) a Japanese face that has not been learned by the memory, and 4) a random pattern. The learned face is reconstructed perfectly by the memory ($r^2 = 1$). The Japanese face is more distorted by the memory than the new Caucasian face ($r^2 = .72$ *vs* $.82$, respectively) and the random vector becomes less and less face-like ($r^2 = .00$).

follows:

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} + \eta(\mathbf{x}_k - \mathbf{W}^{(t)}\mathbf{x}_k)\mathbf{x}_k^T \tag{2.4}$$

where η is a small learning constant and k is randomly chosen. The weight matrix is updated at time $t+1$ by computing the difference between the estimation of the memory $\hat{\mathbf{x}}_k$ (*i.e.,* $\mathbf{W}^{(t)}\mathbf{x}_k$) and the original face vector $\mathbf{x}_k$, and by re-teaching this difference to the memory. This process is repeated many times for every face in the learning set. Smaller and smaller adjustments are made over time until the face is reconstructed perfectly (*i.e.,* until convergence is reached). Figure 3 displays the response of a face memory for a learned face at different steps of Widrow-Hoff learning. The first step (left top panel) corresponds to simple Hebbian learning, the last one (right bottom panel) to complete Widrow-Hoff learning. The quality of reconstruction improves gradually. After complete learning, the face is perfectly reconstructed, and hence, the cosine between the original face and the answer of the memory is equal to one.

In contrast, when a new face is presented as a memory key, it is gradually distorted by the memory in proportion to its similarity to the set of learned faces. New faces that resemble the learned faces are less distorted by the memory than are new faces that are very different from the learned faces. Figure 4 displays the responses of the face memory after complete Widrow-Hoff learning for 1) a learned face, 2) a new face similar to the set of learned faces (Caucasian face), 3) a new face different from the set of learned faces (Japanese face), and 4) a random pattern. The learned face is reconstructed perfectly by the memory ($r^2 = 1$). The new Caucasian face is

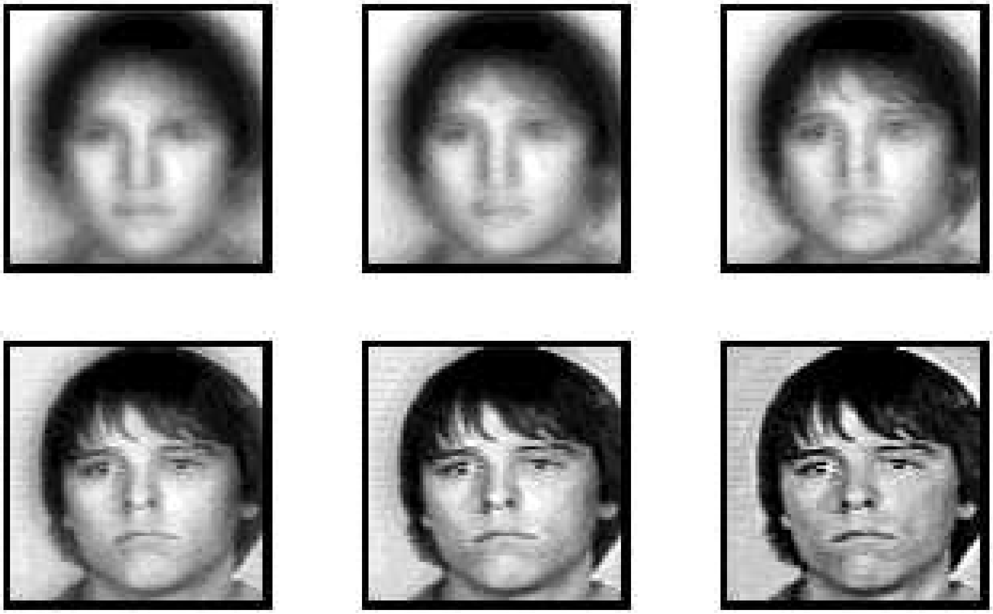

FIG. 3. — Reconstruction of a face at different steps of Widrow-Hoff learning: smaller and smaller adjustments are made until the face is perfectly reconstructed.

the memory is, from left to right: 1) a learned face, 2) a new face similar to the learned face, 3) a new face different from the faces learned by the memory, and 4) a random pattern. Clearly, the memory gives the same response for every stimulus. The fact that this response appears face-like indicates that the memory is able to extract some general "knowledge" about what a face looks like, but is obviously not able to discriminate between faces, or even to distinguish a face from a random pattern. However, the correlation between the original pattern and its reconstruction gives some information about its likelihood of having been stored in the memory. For example, face 1 of Figure 2 was stored in the memory. The squared correlation (r^2) between it and its reconstruction is equal to .71. Face 2 of Figure 2 was not stored in the memory, but comes from the same population as face 1 (*i.e.,* young Caucasians). The r^2 in this case is equal to .66. Face 3 of Figure 2 was not stored in the memory, and comes from a rather different population than face 1 and 2 (*i.e.,* face 1 and 2 are young Caucasians, face 3 is a young Japanese): r^2 equal .10. The random dot pattern of Figure 2 was indeed not stored in the memory and shares no characteristic with a face. Its r^2 is almost zero. Hence, just by using the correlation between the input pattern and its reconstruction, a rather good guess can be made about this pattern having been stored (even though the memory does gives the same response for different cues). Note, however, that the difference in correlation between face 1 and face 2 is rather small which means that the memory is likely to be confused by new patterns similar to the patterns learned.

The performance of the autoassociator can be improved by using a Widrow-Hoff error correction learning rule. The Widrow-Hoff learning rule corrects the difference between the response of the system and the desired response by iteratively changing the weights in $\mathbf{W}$ as

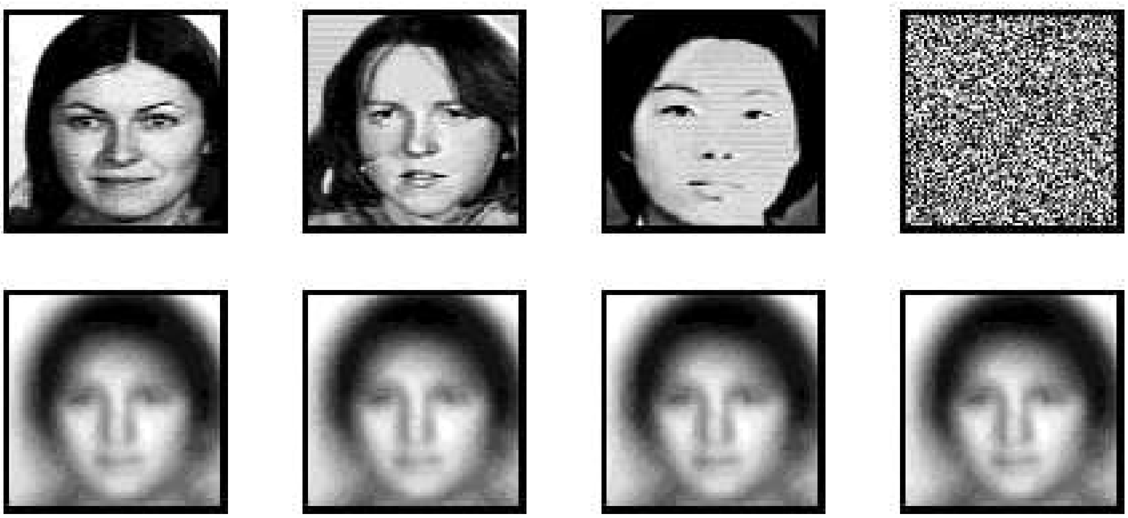

FIG. 2. — The top panels show four stimuli and the bottom panels the "responses" produced by an autoassociative memory trained with 100 Caucasian faces when the stimuli are presented as input. The stimuli are 1) a Caucasian face that has been learned by the memory [$r^2 = .71$], 2) a Caucasian face that has not been learned by the memory [$r^2 = .66$], 3) a Japanese face that has not been learned by the memory [$r^2 = .10$], and 4) a random pattern [$r^2 = .00$]. Clearly, the memory does not discriminate between faces nor even distinguishes a face from a random pattern. The r^2 reported after each stimulus indicates the squared correlation between the stimulus and the answer produced by the memory.

Recall of the k-th face from the memory is achieved by filtering the face through the memory or, more formally, by premultiplying the face vector $\mathbf{x}_k$ by the matrix $\mathbf{W}$:

$$\hat{\mathbf{x}}_k = \mathbf{W}\mathbf{x}_k \tag{2.2}$$

where $\hat{\mathbf{x}}_k$ represents the estimation of the k-th face by the memory. The quality of this estimation can be measured by computing the cosine of the angle between the vectors $\hat{\mathbf{x}}_k$ and $\mathbf{x}_k$, formally:

$$\cos(\hat{\mathbf{x}}_k, \mathbf{x}_k) = \frac{\hat{\mathbf{x}}_k^T \mathbf{x}_k}{||\hat{\mathbf{x}}_k||\ ||\mathbf{x}_k||} \ . \tag{2.3}$$

A cosine of 1 indicates a perfect reconstruction of the stimulus. Another well-known measure of goodness of fit is the coefficient of correlation between the vectors $\hat{\mathbf{x}}_k$ and $\mathbf{x}_k$. The coefficient of correlation noted $r_{\hat{\mathbf{x}}_k,\mathbf{x}_k}$ or simply r, is a particular case of the cosine measure. It is computed as the cosine of the *centered* vectors (*i.e.,* the mean of the vector is subtracted from each component of the vector, so that the transformed vector has zero mean). The square coefficient of correlation (r^2) is, in general, interpreted as the proportion of common variance between two vectors. The comparison between the original and reconstructed images may also be done visually by displaying the memory response.

Figure 2 displays the response of a memory trained with Hebbian learning to store a set of 100 face images (50 Caucasian males and 50 Caucasian females) when the stimulus used to cue

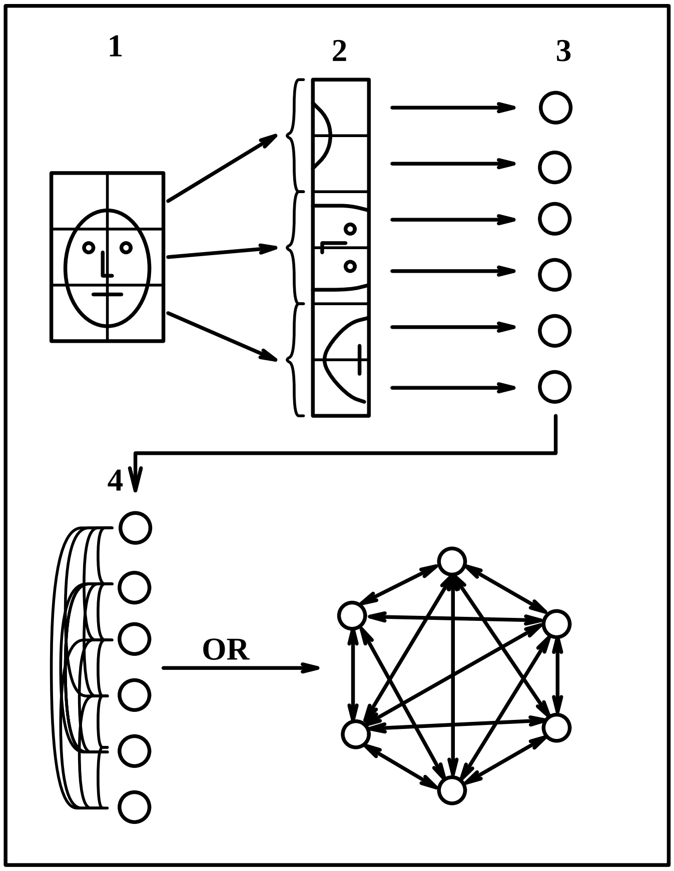

FIG. 1. — Steps in creating an autoassociative memory for face images. *step 1*: the faces are digitized to form a pixel image – *step 2*: the rows of the face images are concatenated to form a column vector – *step 3*: each element of the pixel-vector is used as input to a cell of the autoassociative memory – *step 4*: each cell in the memory is connected to each other cell. Learning is then achieved using either a standard Hebbian learning rule or a Widrow-Hoff error correction learning rule.

Next, each element of the face vector $\mathbf{x}_k$ is used as input to a cell of the autoassociative memory (step 3). The number of cells in the memory, hence, is equal to the number of pixels in the vector $\mathbf{x}_k$. Each cell in the memory is connected to each other cell (step 4). The output of a given cell for a given face is simply the sum of its inputs (*i.e.,* the elements of the face vector, $\mathbf{x}_k$) weighted by the connection strengths between itself and all of the other cells.

The weights or connection strengths are represented by an $I \times I$ matrix $\mathbf{W}$. When standard Hebbian learning is used, $\mathbf{W}$ is obtained by successively autoassociating each face vector $\mathbf{x}_k$, and summing the resultant outer product matrices, with a formula:

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^T \ . \qquad \textbf{(2.1)}$$

To avoid the problem of finding a face representation useful for the wide range of tasks that can be accomplished by human observers, most of the recent statistical/computational models of face processing operate directly on image-based codings of the faces (*i.e.,* 2D arrays of pixel intensities). This type of coding has the advantage of capturing most of the information in faces. In addition to coding implicitly the geometrical information, it preserves detailed featural and textural information. Although models operating on pixel codes of the faces differ in their objective, architecture, algorithm, and particular implementation (*e.g.,* principal component approach [15, 17], linear autoassociative network [1, 8, 10], autoassociative and/or heteroassociative back-propagation network [6, 7, 16]), they all suggest that faces can be represented efficiently in terms of the eigendecomposition of a matrix storing pixel-based descriptions of the faces [18].

In this paper, we present the simplest of these models (*i.e.,* a linear autoassociator) along with its application to the problems of categorizing and identifying human faces. We further discuss its relationship to traditional statistical techniques (*i.e.,* principal component analysis). We wish to demonstrate that when an appropriate coding of the faces is used, a very simple model is able to solve seemingly complex tasks. By the expression "appropriate coding of the faces", we mean a technique that captures the perceptual and statistical properties of the faces. Since the general model of a linear autoassociator has already been described in the first paper of this issue, we shall present only a short description of autoassociative memories, and focus on their application to face images. In the first section, we describe the method used to store and recall face images with a linear autoassociator. Then, we proceed by presenting some appealing properties of face autoassociative memories. Specifically, we show that a linear autoassociator is a useful tool for quantifying/analyzing the perceptual information in face images. Finally, we show that autoassociative memories can be used as a pre-processing device to simulate some psychological tasks, such as categorizing faces according to their sex.

2. Storing faces with a linear autoassociator

Autoassociative memories are a special case of associative memories (cf. Abdi, this issue) in which the input patterns are associated with themselves. The goal of autoassociative memories is to find a set of connections between input units, so that when a portion of an input is presented as a memory key, the memory retrieves the complete pattern, filling in the missing components.

2.1. Model description

The different steps used to store a face image in an autoassociative memory are illustrated in Figure 1. First, the faces are coded as a vector of pixel intensities. This is achieved by digitizing each face to form a pixel image (step 1) and concatenating the rows of the image to form an $I \times 1$ vector $\mathbf{x}_k$ (step 2). Each numeric element in $\mathbf{x}_k$ represents the gray level of the corresponding pixel. In all the simulations presented here, the faces were first digitized from a slide to form a 151×225 pixel image with a resolution of 16 gray levels *per* pixel. Then, a 33975-dimensional vector $\mathbf{x}_k$ was created from each face image by concatenating the rows of the digitized image. For computational convenience the vectors were normalized to unity (*i.e.,* $\mathbf{x}_k^T\mathbf{x}_k = 1$).

Ref: Valentin, D., Abdi, H., O'Toole, A.J. (1994). Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches. *Journal of Biological Systems*, **2(3)**, 413-429.

Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches

DOMINIQUE VALENTIN, HERVÉ ABDI and ALICE J. O'TOOLE

School of Human Development: The University of Texas at Dallas,
MS: GR.4.1., Richardson, TX75083-0688, U.S.A.

ABSTRACT

Recent statistical/neural network models of face processing suggest that faces can be efficiently represented in terms of the eigendecomposition of a matrix storing pixel-based descriptions of a set of face images. The studies presented here support the idea that the information useful for solving seemingly complex tasks such as face categorization or identification can be described using simple linear models (linear autoassociator or principal component analysis) in conjunction with a pixel-based coding of the faces.

Keywords: face processing, autoassociative memory, principal component analysis, macro features.

1. Introduction

The apparent ease with which human observers recognize faces should not mask the complexity of the operations involved in this processing. The human face is a complex, multidimensional, and meaningful pattern that poses an exciting challenge for computational modeling. Faces are highly similar to one another, containing the same features arranged in roughly the same configuration. Yet, human observers are able to discriminate, remember, and identify an impressive number of faces across large changes in pose, expression, lighting and so on. To simulate such performance, computational models should be able to encode very subtle variations in the form and configuration of facial features, which are difficult to describe in traditional feature-based terms.