Diamantaras, K.I., & Kung, S.Y. (1996). *Principal component neural networks: Theory and applications*. New York: J. Wiley

Etcoff, N.L., Freeman, R., & Cave, K.R. (1991). Can we lose memories for faces? Content specificity and awareness in a prosopagnosic. *Journal of Cognitive Neuroscience, 3*, 25–41.

Ginsburg, A.P. (1978). Visual information processing based on spatial filters constrained by biological data. *AMRL Technical Report*, 78–129, Ohio.

Hancock, P.J., Burton, M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory and Cognition, 24*, 26–40.

Harmon, L. (1973). The recognition of faces. *Scientific American, 229*, 71–82.

Harmon, L.D. & Hunt, W.K. (1977). Automatic recognition of human face profiles. *Computer, Graphics and Image Processing, 6*, 135–156.

Harmon, L.D., Khan, M.K., Lash, R., & Ramig, P.F. (1981). Machine identification of human faces. *Pattern Recognition, 13*, 97–110.

Horn, R.A & Johnson, C.R. (1985). *Matrix Analysis*. Cambridge: C.U.P.

Kaya, Y. & Kobayashi, K. (1972). A basic study on human face recognition. *International Conference on Frontiers of Pattern Recognition*, 265–289.

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

McNeil, J.E. & Warrington, E.K. (1991). Prosopagnosia: A reclassification. *The Quarterly Journal of Experimental Psychology, 43A*, 267–287.

Nakamura, O., Mathur, S., & Minami, T. (1991). Identification of human faces based on isodensity maps. *Pattern Recognition, 24*, 263–271.

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., Abdi, H., Deffenbacher, K.A., & Bartlett, J.C. (1991). Classifying faces by race and sex using an autoassociative memory trained for recognition. In K.J. Hammomd, & D. Gentner (Eds.), *Proceedings of the thirteenth annual conference of the cognitive science society*. Hillsdale, N. J.: Erlbaum.

O'Toole, A.J., Abdi, H., Deffenbacher, K.A., & Valentin, D. (1993). A low-dimensional representation of faces in the higher dimensions of the space. *Journal of the Optical Society of America A, 10*, 405–411.

O'Toole, A.J., Abdi, H., Deffenbacher, K.A., & Valentin, D. (1995). A perceptual learning theory of the information in faces. In T. Valentine (Ed.), *Cognitive and computational aspects of face recognition*. Routledge: London.

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

O'Toole, A.J., Peterson, J., & Deffenbacher, K.A. (1996). An other-race effect for classifying faces by sex. *Perception, 25*, 669–675.

O'Toole, A.J., Vetter, T., Troje, N.F., & Bülthoff, H.H. (1997). Sex classification is better with three-dimensional head structures than with image intensity information. *Perception, 26*, 75–84.

Sakai, T., Nagao M., & Kidode, M., (1971). Processing of multilevel pictures by computer - the case of photographs of human face. *System, Computers,& Controls, 2*, 47–54.

Samal, A. (1991). Minimum resolution for human face detection and identification. *proceedings of SPIE human vision, visual processing, and digital display, 1453*, 81–89.

Samal, A. & Iyengar, P.A. (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition, 25*, 65–77.

Sergent, J. (1986a). Microgenesis of face perception. In H.D. Ellis, M.A. Jeeves, F. Newcombe, & A. Young (Eds.) *Aspects of face processing*. Dortrecht: Martinus Nijhoff.

Sergent, J. (1986b). Methodological constraints on Neuropsychological studies of face perception in normals. In R. Bruyer (Ed.) *The Neuropsychology of face perception and facial expression*. Hillsdale: Lawrence Erlbaum.

Sirovich, L. & Kirby M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A, 4*, 519–524.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive NeuroSciences, 3*, 71–86.

Valentin, D., Abdi, H., & O'Toole, A.J., & Cottrell, W.G. (1994). Connectionist models of face processing: A survey. *Pattern recognition, 27*, 1209–1230.

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*, 413–429.

Valentin, D. & Abdi H. (1996). Can a linear autoassociator recognize faces from new orientation? *Journal of the Optical Society of America, A, 13*, 717–724.

Valentin, D., Abdi, H., Edelman, B., & Nijdam, A. (1996). Connectionism "face"-off: Different algorithms for different tasks. *Psychologica Belgica, 36*, 65-92.

Valentin, D., Abdi, H., & Edelman, B. (in press, 1997). What represents a face: A computational approach for the integration of physiological and psychological data. *Perception*,

Young, A.W. & Ellis, H.D. (1989). Childhood Prosopagnosia. *Brain and Cognition, 9*, 16–47.

Young, A.W., Hellawell, D., & De Haan, E.H.F. (1988). Cross-domain semantic priming in normal subjects and a prosopagnosic patient. *Quarterly Journal of Experimental Psychology, 40*, 561–580.

Yuille, A.L. (1991). Deformable templates for face recognition. *Journal of Cognitive Neurosciences, 3*, 59–70.

identity specific information will be impaired. The *categorical* information is more resilient to degradation and hence is more likely to be preserved in the case of a local as well as a distributed lesion.

Such a selective impairment (*i.e.,* destruction of identity information with preservation of categorical information) is reminiscent of some neuropsychological data. In several cases of prosopagnosia (*i.e.,* inability to recognize faces), a simple dissociation between identity and categorical information has been reported (*e.g.,* Etcoff, Freeman & Cave, 1991; McNeil & Warrington, 1991; Young & Ellis, 1989). In general, patients exhibiting this dissociation are unable to identify a face or to decide whether they know the face, but are able to decide whether it is a female or male face, or what is the approximate age of the person. This is consistent with the fact that information related to the identity of faces is less robust than the categorical information. Additionally, in accordance with the idea that random lesions of a face autoassociative memory would not selectively impair the information conveyed by the eigenvectors with large eigenvalues, double dissociations between identity and categorical information for faces have not (to the best of our knowledge) been reported in the literature. In other words, cases of patients able to identify or recognize faces, but not able to derive the gender of a person (or other "semantic" information, with the exception of emotion) from his/her face have not been reported.

However, as mentioned by Burton, Young, Bruce, Johnston, and Ellis (1991), it is important to note that face recognition deficits might occur at different levels of processing (*e.g.,* perceptual, semantic, episodic). The analogy drawn between the autoassociative memory lesioning and prosopagnosic patients applies only to the case of perceptual deficits. Further, if we were to extend this analogy to more complex phenomena such as *covert face recognition*, an additional decision mechanism would be necessary. Covert recognition refers to the fact that some prosopagnosic patients with no overt recognition of faces (explicit memory) actually demonstrate some recognition ability when indirectly tested (implicit memory). For example, Young, Hellawell, and De Haan (1988) reported the case of a prosopagnosic patient (PH) who was unable to retrieve identity of familiar faces but showed normal recognition effects when tested covertly. To give an account of covert recognition in terms of the PCA approach one would have to assume that part of the information contained in the eigenvectors with relatively small eigenvalues is preserved. This information would then be used by some kind of a decision system, the role of which would be to evaluate the amount of preserved information. Such a decision system could involve two different thresholds, a "recognition threshold" and an "explicit recognition threshold". Covert recognition would occur if the amount of information provided by the eigenvectors with small eigenvalues lies between those two thresholds. But clearly, such a mechanism is quite speculative at this time and further study is necessary to determine its psychological relevance.

In conclusion, the internal representation extracted by an autoassociative memory from a set of faces seems to share some properties with the information human observers extract from faces (see, also Valentin, Abdi, Edelman, in press, for an elaboration on this point of view). However, to assess the psychological relevance of this type of facial representation, future research needs to explore more precisely the relationship between human performance on face processing tasks and the prediction of the autoassociative memory for the same tasks.

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. (1994a). *Les réseaux de neurones.* Grenoble: Presses Universitaires de Grenoble.

Abdi, H. (1994b). A neural network primer. *Journal of Biological System, 2,* 247–281.

Abdi, H., Valentin, D., Edelman B., & O'Toole, J.A. (1995). More about the difference between men and women: Evidence from linear neural networks and the principal component approach. *Perception, 24.* 539–562.

Anderson, J.A. and Mozer, M. (1981). Categorization and selective neurons. In G. Hinton and J. Anderson (Eds.) *Parallel models of associative memory.* Hillsdale, N.J.: Erlbaum.

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., Ellis, H., Gibling, F., & Young, A.W. (1987). Parallel processing of the sex and familiarity of faces. *Canadian Journal of Psychology, 41,* 510–520.

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

Burton, M. Young, A.W. Bruce, V., Johnston, R.A., & Ellis, A.W. (1991). Understanding covert recognition. *Cognition, 39,* 129–166.

Cottrell, G.W. & Fleming, M. (1990). Face recognition using unsupervised feature extraction. *Proceedings of the International Conference on Neural Network* pp. 322–325.

Eigenvectors	Number of faces per sample								
	2	4	6	8	10	20	50	100	150
1	.94	.97	.98	.99	.99	.99	.99	.99	.99
2	.39	.47	.54	.56	.58	.69	.84	.96	.99
3		.28	.33	.39	.42	.58	.81	.94	.99
4		.15	.23	.28	.29	.51	.80	.94	.99
5			.15	.23	.24	.39	.67	.92	.98
6				.14	.22	.30	.53	.86	.98
7				.14	.18	.25	.48	.83	.98
8					.13	.20	.43	.85	.95
9					.11	.20	.37	.73	.93
10						.17	.27	.58	.79
20							.15	.29	.49
50								.12	.30
100									.16

TABLE 3. Average correlation between the eigenvectors extracted from random samples of faces and the eigenvectors extracted from the complete set of faces as a function of the number of faces per sample.

In summary, the pattern of stability of the eigenvectors, reflected in the data of Table 3, is the following: very high stability of the first eigenvector, lesser but still good stability of the next five eigenvectors and decreasing stability of the eigenvectors with smaller eigenvalues. This differential degree of availability and robustness of the information conveyed by different eigenvectors can be related to some temporal properties of the visual system reported by Sergent (1986a). Briefly stated, some physiological and psychophysical evidence suggests that the visual system does not instantaneously nor simultaneously extract all the information available in a visual stimulus. The quantity of information available increases as a function of the exposure duration and decreases with the retinal eccentricity. This "microgenesis" of perception suggests that different kinds of information, conveyed by different spatial frequencies, are processed at different speeds. Particularly, since low spatial frequencies are resolved faster than high spatial frequencies, information relative to categorical discrimination is available earlier than information relative to the identification of a face. In agreement with that idea, empirical evidence indicates that gender decisions are always made much faster than identity (*i.e.*, familiar versus unfamiliar) decisions, suggesting that computation of gender and identity might be two independent processes (Bruce, Ellis, Gibling & Young, 1987).

4. CONCLUSION

The results presented here support the idea that different kinds of facial information are conveyed by different ranges of eigenvectors of a cross-product matrix derived from a set of face images. These different types of information have different properties and are not equally useful, depending on the type of tasks to be performed. The eigenvectors with large eigenvalues contain mostly low frequency information, are very stable, and capture information that is generalizable to new faces. In contrast, eigenvectors with small eigenvalues contain essentially high frequency information, are very unstable and, the information they capture is not generalizable to new faces.

An interesting implication of the pattern of stability found in the third series of simulations, is that the information conveyed by the eigenvectors with relatively small eigenvalues is the most vulnerable to any kind of degradation. Hence, if we were to randomly lesion an autoassociative memory trained to reconstruct a set of face images, only the information conveyed by the eigenvectors with small eigenvalues would be degraded substantially. In other words, only

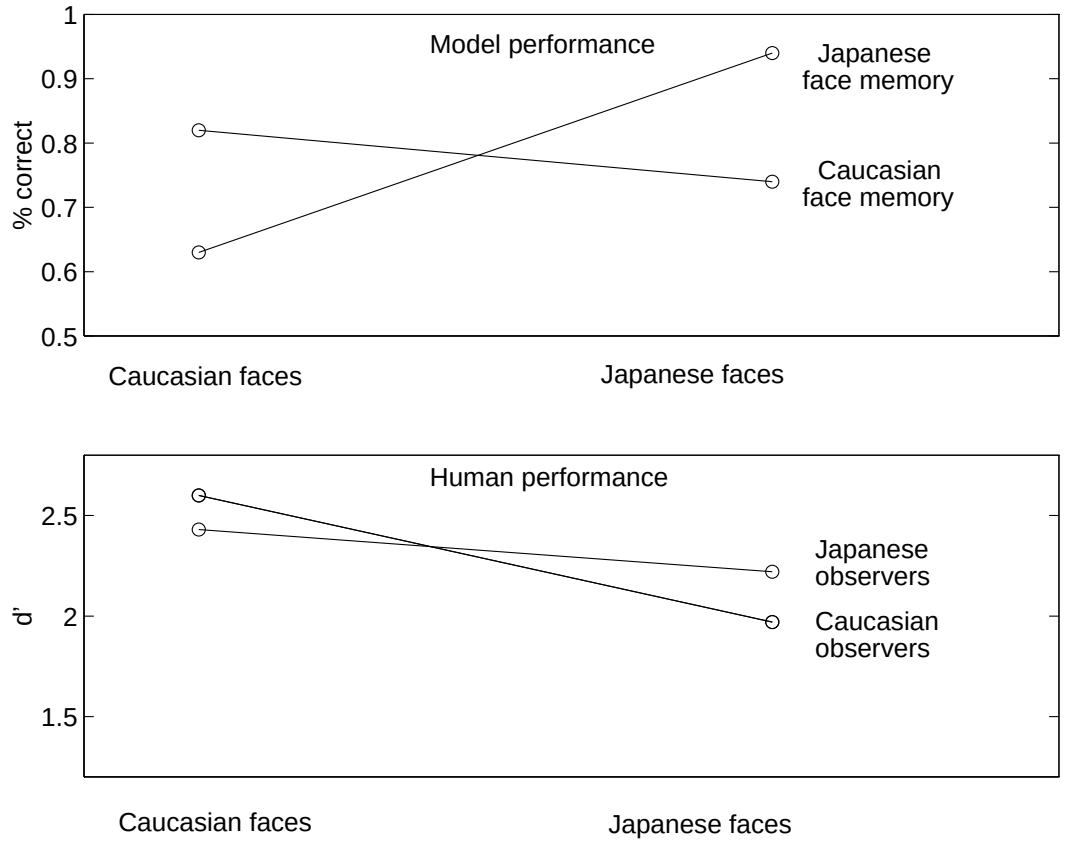

FIGURE 10. (a) Proportion of correct gender classifications for new Caucasian and Japanese faces by a Caucasian and a Japanese autoassociative memory (top panel); (b) accuracy of gender categorization (d') for Caucasian and Japanese faces by Caucasian and Japanese observers (bottom panel).

captures the information that is most common to all faces and because, as mentioned previously, the first eigenvalue is very large (81%) this eigenvector should be the most stable. What is impressive, however, is that only 2 faces are necessary to estimate it.

According to O'Toole *et al.* (1991, 1993), the information useful for categorizing faces by gender, race, and probably age, is conveyed essentially by the eigenvectors with the larger eigenvalues. With the particular set of faces used in this simulation, the 2nd, 3rd, and 4th eigenvectors convey most of the information related to the general sexual appearance of the faces. Table 3 shows that, on the average, 50 faces (25 males, 25 females) are sufficient to estimate about 65% of these eigenvectors ($r = .80$). With 100 faces (50 males, 50 females), about 90% of the same eigenvectors are correctly reconstructed ($r = .94$).

For the next five eigenvectors (*i.e.*, up to 9), the accuracy of reconstruction decreases progressively from 84% ($r = .92$) to 53% ($r = .73$) with 100 faces. With 150 faces, the accuracy of reconstruction of the same eigenvectors decreases from 96% ($r = .98$) to 86% ($r = .93$). The quality of reconstruction of the eigenvectors with smallest eigenvalues decreases drastically from 62% for the 10th to 3% for the 100th eigenvector with a sample of 150 faces. Again, this is not surprising because the eigenvectors with relatively small eigenvalues convey essentially information specific to individual or small groups of faces. These eigenvectors are specific to the particular sample learned.

FIGURE 9. The first two eigenvectors of an autoassociative memory created from 25 male and 25 female face images.

short hair" is highly efficient to dissociate male and female faces for this particular set of Japanese faces, it not a very useful information to dissociate male and female faces for the set of Caucasian faces. This result indicates that the generalizability of categorical information depends in part on the homogeneity of the set of faces from which it is extracted.

3.4. **Estimation of eigenvectors stability.** The purpose of this third series of simulations was to estimate the stability of the information conveyed by the eigenvectors. Specifically, we were interested in finding the minimum number of faces necessary to estimate different eigenvectors accurately.

3.4.1. *Procedure.* An autoassociative memory was created using the complete set of faces, and decomposed into its eigenvectors. To assess the stability of these eigenvectors, face samples, ranging in size from 2 to 150, were randomly selected from the original set of faces, with the constraint that each sample contained half male and half female faces. For each sample, an autoassociative memory was created and decomposed into its eigenvectors. Then, a coefficient of correlation between the eigenvectors extracted from each sample and the eigenvectors extracted from the full set of faces was computed. This process was repeated 100 times, for each sample size, to ensure that the results were not sample-dependent.

3.4.2. *Results and Discussion.* The average correlations are reported in Table 3. Note that the eigenvectors are ordered according to their eigenvalues. As previously noted, 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. A simple glance at Table 3 indicates that the minimum number of faces necessary to estimate correctly the original eigenvectors varies as an inverse function of their eigenvalues. The larger the eigenvalue associated with an eigenvector, the fewer faces needed to estimate the eigenvector.

Examination of the first line of Table 3 shows that, on the average, only two faces are sufficient to estimate 88% of the original first eigenvector ($r = .94$, $r^2 = .88$), and 8 faces are necessary to estimate it with 98% accuracy ($r = .99, r^2 = .98$). This indicates that the information contained in this eigenvector is very robust and easily accessible. This is not really surprising since, by definition, the first eigenvector

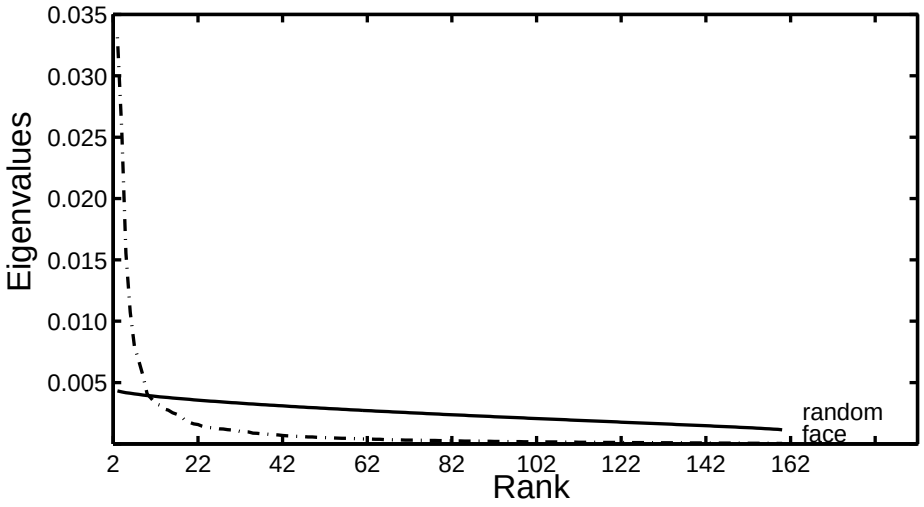

FIGURE 8. Randomization test for the eigenvalues: The solid line represents the eigenvalues obtained after randomization of the face pixel matrix **X**. The dashed line represents the eigenvalue of the face autoassociative memory **W**. For readability purposes, the eigenvalue corresponding to the first eigenvector was dropped.

difference in the patterns of results depicted in Figure 10a and b is a reversal of main effect. For the simulations, we obtained a main effect of face race, with Japanese faces yielding better performance than Caucasian faces. A main effect of face race was also obtained by O'Toole *et al.* but in favor of Caucasian faces. This difference is probably due to the fact that in O'Toole *et al.* the faces were cropped so as to eliminate hair information. This contention seemed to be confirmed by the fact that in a pilot study with the same faces, including the hair, O'Toole *et al.* found that the gender of Japanese faces was identified more quickly than that of Caucasian faces.

The superiority of hairy Japanese faces over hairy Caucasian faces can be explained by a larger homogeneity within the subsets of male and female Japanese faces than within the subsets of male and female Caucasian faces. Since this superiority seems to vanish when the hair information is discarded, we can speculate that this effect was due essentially to the fact that, for our face samples, Japanese hair-do are more sexually stereotyped than Caucasian hair-do. An inspection of the face images, indeed, showed that all Japanese males have very short hair, whereas most Japanese female have mid-length to long hair. This obvious distinction does not exist for the Caucasian images where some males have mid-length hair and some females very short hair. However, further work would be neccessary to clarify this issue.

A last point worth noting on this simulation series is the difference in generalization power of the two autoassociative memories. The Japanese autoassociative memory generalizes better to own race faces than the Caucasian autoassociative memory (94% *vs.* 82%) but not to other race faces (63% *vs.* 74%). Again this difficulty in generalizing to other race faces is probably due to the sexually stereotyped Japanese hair-do. Because of the large within subsets homonegeity, the information captured by eigenvectors with large eigenvalues is highly efficient to categorize faces coming from these subsets, but less so for other faces. For example, while the information "having a very

	$\mathbf{W}_{\text{rand}}$			$\mathbf{W}_{\text{face}}$
Rank	Average	Minimum	Maximum	Average
1	.6075	.6072	.6079	.8137
2	.0043	.0042	.0045	.0331
3	.0042	.0041	.0044	.0255
4	.0042	.0041	.0043	.0156
5	.0041	.0040	.0042	.0107
6	.0041	.0040	.0042	.0080
7	.0040	.0039	.0041	.0068
8	.0040	.0039	.0041	.0056
9	.0039	.0039	.0040	.0041
10	.0039	.0038	.0040	.0037

TABLE 2. Randomization test for the distribution of the eigenvalues of the face autoassociative memory under the null hypothesis of absence of structure. The first three columns represent the average, maximum, and minimum proportion of variance explained for the eigenvectors of $\mathbf{W}_{\text{rand}}$ and the last column the average proportion of variance for the eigenvectors of the face matrix $\mathbf{W}$.

Caucasian face images and twenty samples of 50 Japanese face images were used as training sets. The remaining faces (Caucasian and Japanese) were used to test the ability of the memory to generalize to new own-race and other-race faces. The estimation of the gender of the faces was performed using the algorithm described previously, with the difference that, because we wanted to evaluate the optimal performance of the memories, all the eigenvectors were used for this series of simulations.

3.3.2. *Results and discussion.* Figure 10a presents the proportion of correct gender classifications for new Caucasian and Japanese faces by the Caucasian and Japanese autoassociative memories. Three major points can be noted from this figure:

- Both the Caucasian and the Japanese autoassociative memories perform better with own-race faces (82% and 94%, respectively) than with other-race faces (63% and 74%, respectively).
- Japanese faces, however, seem to be on the whole easier to gender categorize than Caucasian faces (84% *vs* 72%).
- All performance, however, is above chance level, thus indicating that gender information is partially generalizable across races.

In summary, this series of simulations showed that part of the gender information captured by the eigenvectors of a population of faces can be generalized to another population. This suggests that gender information is partially common to all faces independent of the race of the faces. However, the decrement in performance between own and other race faces also suggests that it is easier to process gender information in the context of a familiar race than in the context of a different race. This result is interesting in that such an other-race effect for gender categorization has been described for human subjects. Using the same face database as here, O'Toole, Peterson and Deffenbacher (1996), found Oriental observers to be more efficient than Caucasian Observers at gender categorizing Japanese faces, and *vice versa*.

As a comparison point between simulations and human data, Figure 10b presents the d' obtained by Caucasian and Oriental observers for Caucasian and Japanese faces in the O'Toole *et al.* study. The main

tests. The procedure used was as follows. The elements of matrix $\mathbf{X}$ were randomly permuted to give the matrix $\mathbf{X}_{\text{rand}}$, and the eigenvalues of the matrix $\mathbf{W}_{\text{rand}}$ (*i.e.,* $\mathbf{X}_{\text{rand}}\mathbf{X}_{\text{rand}}^T$) were then computed. The procedure was repeated 2000 times in order to be able to derive an empirical sampling distribution of the eigenvalues of $\mathbf{W}$ under the null hypothesis of absence of structure in the matrix $\mathbf{X}$. The average value of the first 10 eigenvalues of $\mathbf{W}_{\text{rand}}$ as well as their minimum and maximum values are reported in Table 2 with the corresponding eigenvalues of $\mathbf{W}$. The corresponding values for the second eigenvalues onward are plotted in Figure 8 (the first eigenvalue is omitted because its size makes the plot of the other eigenvalues unreadable).

The relatively high values of the first random eigenvalue are due to the fact that $\mathbf{W}$ is a *cross-product* matrix (instead of, say, a covariance matrix), and hence the first eigenvector is almost equivalent to an average. In the case of the face autoassociative memory, the first eigenvector is very similar to the average face[3] (*i.e.,* the coefficient of correlation between the average face and the first eigenvector is $r = .9996$). However, the eigenvalue of the face autoassociative memory (81%) is very clearly larger than what can be expected under the null hypothesis (61%). Therefore, the importance of the eigenvalue associated with the first eigenvector reflects not only the fact that $\mathbf{W}$ is a cross-product matrix but also the very strong intersimilarity of faces. The specific structure of the set of faces is also reflected by the difference in the distribution of the face eigenvalues from the test eigenvalues (cf. Figure 8). Clearly, while in the test condition all the eigenvalues after the first one are almost equivalent (*i.e.,* the space is almost spherical), in the face condition, the eigenvalues decrease progressively.

As an illustration of the perceptual information captured by earlier eigenvectors, Figure 9 displays the first two eigenvectors of an autoassociative memory created from 25 female and 25 male faces. The first eigenvector appears clearly to be some kind of an average face. The second eigenvector is somewhat harder to describe. At first sight, the hair area seems to be an important contributor to this eigenvector. This suggests that, for the sample used in the present simulation, the hair might be an important factor for gender discrimination. However, a more precise analysis of the pixel contributions to this eigenvector, showed that the forehead, eyebrows, nose, and chin areas also contribute strongly to this eigenvector. On the average, male faces in this sample tend to have a longer chin, a bigger nose, thicker eyebrows, and shorter hair than female faces (Abdi *et al.*, 1995).

Finally, an analysis of the errors achieved by the model shows that the system tends to be biased to classify faces as male. This bias can be explained by the fact that female faces are more dispersed around their barycenter than male faces (cf. Figure 7). Hence, some female faces are closer to the barycenter of the male faces than they are to their own barycenter. This bias could be easily avoided by using a classification algorithm that takes into account the unequal dispersion of the female and male faces around their barycenters.

3.3. Other race effect in gender categorization.

The purpose of this second series of simulations was to determine the extent to which the gender information extracted from a sample of faces can be generalized to faces from a different population. Previous work showed that the eigenvectors of an autoassociative memory created from a heterogeneous set of faces, made up of a majority of faces of one race and a minority of faces of another race, are optimal for discriminating between majority-race faces but not between minority-race faces (O'Toole, Abdi *et al.*, 1991). To determine whether the PCA approach would predict such an other-race effect for gender categorization, we compared the ability of a Caucasian and a Japanese autoassociative memory to classify new Caucasian and Japanese faces. If the gender information contained in face images is independent of the race of the face, or in other words, if this information is generalizable across race, the performance for Caucasian and Japanese faces should be roughly equivalent. On the other hand, if gender information depends on the race of the faces, the Caucasian autoassociative memory should classify better new Caucasian faces than new Japanese faces and *vice versa.*

3.3.1. *Procedure.* The procedure was similar to that used in the first simulation. Twenty samples of 50

[3]The orthogonality of the eigenvectors makes the extraction of the first eigenvector serve as a sort of centering process (*i.e.,* subtracting the average face from each face), and hence the $(\ell + 1)$th eigenvector of $\mathbf{W}$ is very similar to the ℓth eigenvector of a face covariance matrix (like the one used, for example, by Turk & Pentland, 1991).

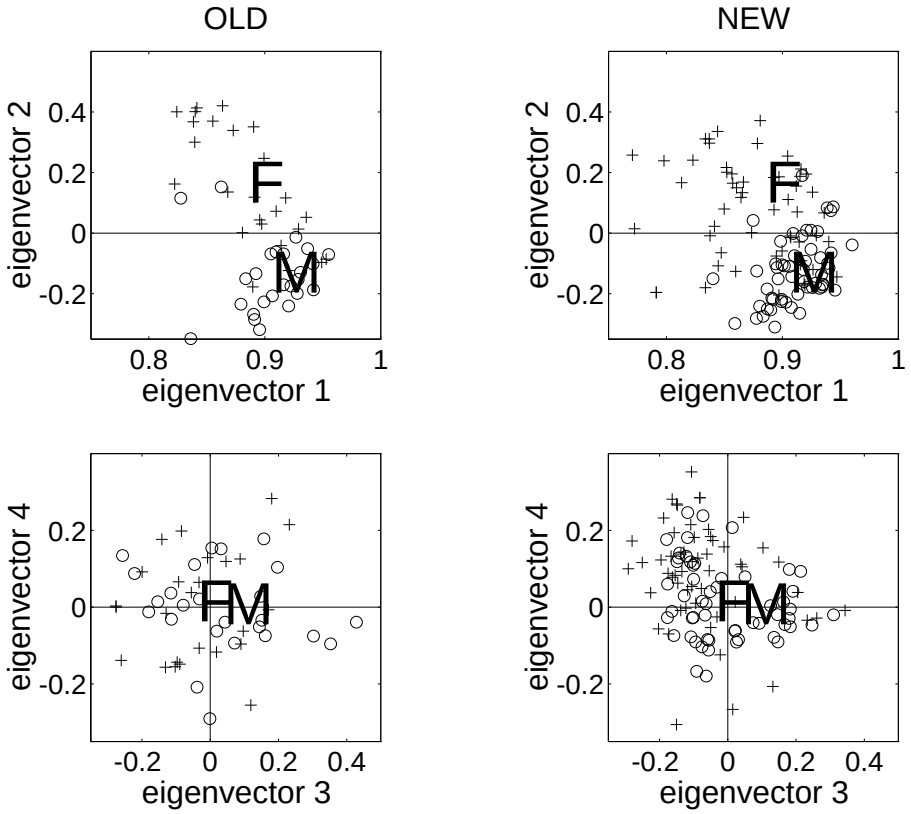

FIGURE 7. Projections of the 160 Caucasian faces in the database onto the first four eigenvectors of a sample of 50 faces. The projections of the 50 training faces ("old") appear on the left panels and the projections of the 110 new faces on the right panels. The barycenters of the male ("o") and female ("+") faces are represented respectively by the letters "**M**" and "**F**". The second eigenvector is the most important one for discriminating between male and female faces.

Eigenvectors	1	2	3	4
r	.28	.68	.24	.03
r^2	.08	.47	.05	.00

TABLE 1. Correlation analysis between the projections of the faces onto the first 4 eigenvectors and the gender of the faces coded as "0" and "1".

The first eigenvector represents essentially the characteristics shared by all the faces or, in other words, it represents the face category in general. Because both old (left panel) and new (right panel) faces are strongly correlated with this eigenvector (the average squared correlation between a face and the first eigenvector is .8), it could be used to categorize faces as opposed to other object categories or to detect a face in an image. The eigenvalue associated with this eigenvector is very large: it represents 81% of the total of the eigenvalues (*i.e.*, it explains 81% of the total variance). To evaluate the importance of this value, we performed a series of randomization

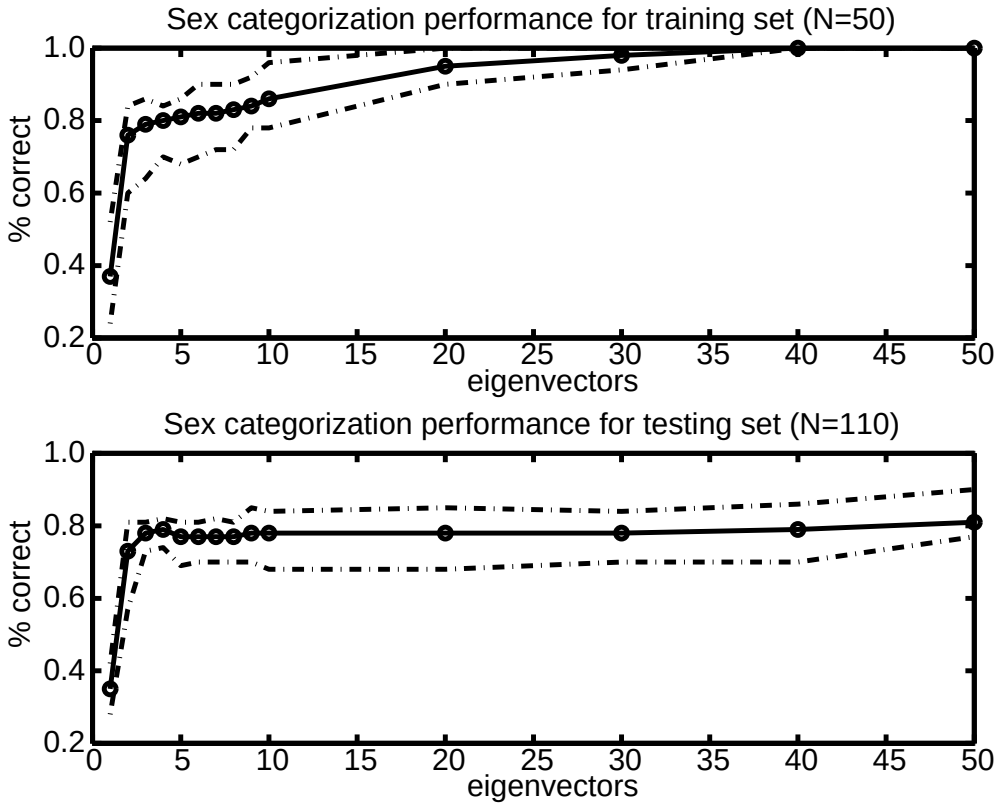

FIGURE 6. Average, minimum, and maximum proportion of correct classification for learned ("old") and unlearned ("new") faces (collapsed across gender) as a function of the number of eigenvectors used to reconstruct the faces.

al. (1995) found that with a training set of 158 faces, about 90% of the new faces were correctly classified as male or female. In other words, going from a training set of 50 faces to a training set of 158 faces increases the performance by less than 10%. This result, in conjunction with the small performance dispersion displayed in Figure 6, suggests that generalizable gender information can be derived from a small sample of faces and remains stable from one sample to the other. In summary, this series of simulations corroborates the Abdi *et al.* finding that all eigenvectors contain information relative to the gender of faces, but that only the information conveyed by eigenvectors with large eigenvalues can be generalized to new faces. In addition, it shows that these eigenvectors can be estimated from a relatively small number of faces. We shall investigate this issue further in the third series of simulations.

As an illustration of the role of the first four eigenvectors in gender classifying faces, Figure 7 displays the projections of the faces used in the first series of simulations onto the first four eigenvectors of a sample of 50 training faces. The training faces are represented on the left panels and the new faces on the right panels. Observation of the distance between male **M** and female **F** barycenters shows that the second eigenvector is the most reliable gender predictor. This observation was confirmed by a correlation analysis between the projections of the faces onto each eigenvector and their gender (cf. Table 1).

50 faces each. This was aimed to assess the robustness of the type of face representation proposed by the PCA approach.

3.2.1. *Procedure.* Different samples of 50 Caucasian faces (25 males and 25 females) randomly selected from the original set of 160 Caucasian face images were used as training sets. The remaining faces were used to test the ability of the memory to generalize to new faces. The faces in the training sets were represented by $I \times 50$ matrices denoted $\mathbf{X}_{\text{old}}$ and the faces in the training sets by $I \times 110$ matrices denoted $\mathbf{X}_{\text{new}}$. The estimation of the gender of the faces by the model was performed using the following algorithm.

1. For each training set, an autoassociative memory (cf. Equation 5) was created from the face images and decomposed into its eigenvectors:
$$\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \ . \tag{10}$$
2. The projections of all the faces (learned and new) onto the first N (with $N < L$) eigenvectors (*i.e.,* the ones with the largest eigenvalues) were computed as:
$$\mathbf{G}_{\text{old}} = \mathbf{X}_{\text{old}}^T \mathbf{U}\mathbf{\Delta}^{-1} = \mathbf{V} \tag{11}$$
for the learned faces, and
$$\mathbf{G}_{\text{new}} = \mathbf{X}_{\text{new}}^T \mathbf{U}\mathbf{\Delta}^{-1} \tag{12}$$
for the new faces. Recall that, after complete Widrow-Hoff learning, the variance of the projections onto each eigenvector is equal to 1 (*i.e.,* the weight matrix is sphericized). This is done by multiplying $\mathbf{U}$, the eigenvectors of $\mathbf{W}$, by $\mathbf{\Delta}^{-1}$ in Equations 11 and 12.
3. The coordinate vectors of the average male face ($\mathbf{m}$) and the average female face ($\mathbf{f}$) were computed by taking the mean of the projections of the male and female learned faces onto the first N eigenvectors, respectively:
$$\mathbf{m} = \frac{1}{J} \sum_{j \in \{\text{male faces}\}}^{J} \mathbf{g}_j \tag{13}$$
and
$$\mathbf{f} = \frac{1}{J'} \sum_{j' \in \{\text{female faces}\}}^{J'} \mathbf{g}_{j'} \tag{14}$$
where J represents the number of learned male faces, J' the number of learned female faces; and $\mathbf{g}_j$ and $\mathbf{g}_{j'}$ the projection of the jth and j'-th face, respectively, onto the N first eigenvectors.
4. The categorization of a face was determined on the basis of the Euclidean distance between its projections onto the first N eigenvectors and the coordinate vectors (in N dimensions) of the average faces. Specifically, the distance of the kth face to the average male and female faces in the N-dimensional subspace is computed as
$$d(\mathbf{g}_k, \mathbf{m}) = ||\mathbf{g}_k - \mathbf{m}|| \tag{15}$$
and
$$d(\mathbf{g}_k, \mathbf{f}) = ||\mathbf{g}_k - \mathbf{f}|| \ . \tag{16}$$
Faces closer to the average female face were classified as female, and faces closer to the average male face were classified as male.

Note that in a neural network framework this is equivalent to using a simple perceptron with the projections of the faces onto the first N eigenvectors as input, and the gender of the faces as output (cf. Abdi, 1994b). The number of eigenvectors (N) used to perform the categorization task varied from 1 to 50 (50 represents the rank of the matrix $\mathbf{W}$ and hence gives the maximum number of eigenvectors). For each condition (*i.e.,* number of eigenvectors), 20 different samples were randomly selected from the original set of faces.

3.2.2. *Results and Discussion.* Figure 6 presents the average (solid lines) and the minimum and maximum (dashed lines) proportions of correct classifications for old and new faces. The results are collapsed across gender. This figure shows that, for the old faces (top panel), the accuracy of categorization increases with the number of eigenvectors used to reconstruct the faces. When only the first eigenvector is used, performance is below the chance level (37% correct classification averaged across samples). When the second eigenvector is added to the first one, performance improves dramatically (76%). When more eigenvectors are used, performance increases smoothly until a perfect categorization score is obtained with 40 eigenvectors.

For the new faces, the accuracy of categorization increases significantly with the first four eigenvectors and then reaches a plateau with a value of 83% correct classification (plus or minus 5% depending on the sample). Using the same face database, Abdi *et*

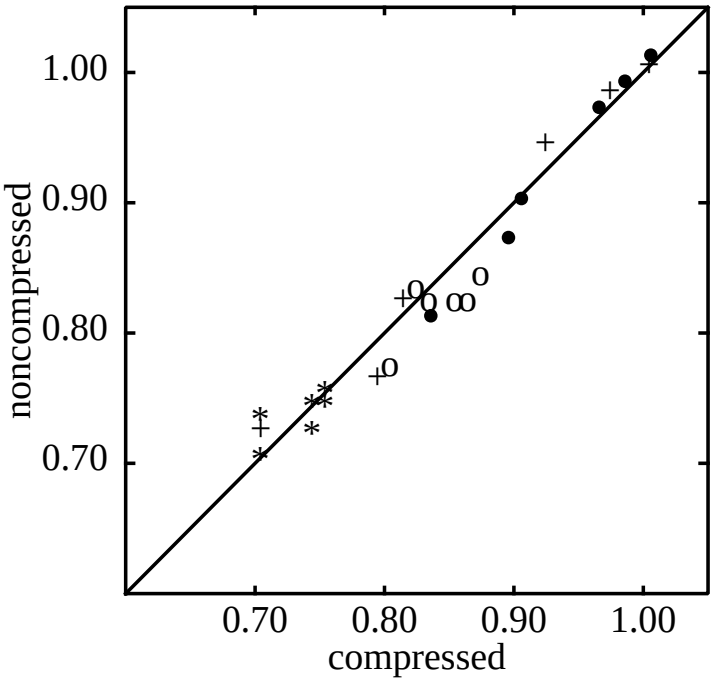

FIGURE 5. Proportion of correct classifications obtained with compressed faces versus proportion of correct classifications obtained with non compressed faces. $*$: new female faces; $+$: old female faces; $\circ$: new male faces, $\bullet$: old male faces. The training sets were composed of 50 faces (25 males, 25 females). The number of eigenvectors used to reconstruct the faces were 2, 5, 10, 20, 30, 40. For each condition, the proportion of correct classifications was averaged across 20 random samples.

faces against the proportion of correct classifications obtained with compressed faces.

3.2. **Gender Classification.** The purpose of this first series of simulations was to replicate and expand the results of Abdi *et al.* (1995). Abdi *et al.* trained two classification networks (perceptron and radial basis function network) to classify a set of face images according to their gender. The face images were either preprocessed via an eigendecomposition or used directly as input to the classification networks. They showed that the eigendecomposition preprocessing not only saves processing time by reducing the size of the classification networks, but also produces a set of features relevant for discriminating between male and female faces. In addition, they showed that although all eigenvectors contain information about the gender of a given face, only the information captured by the eigenvectors with large eigenvalues is useful to classify new faces. These results were obtained using a leave-one-out jackknife technique that maximizes the number of training faces. In the simulations reported here, a bootstrap technique was used to test different random samples of

FIGURE 4. The top panels show a Japanese face reconstructed using 1) the first 40 eigenvectors of an autoassociative memory trained with 160 Caucasian faces ($r = .87$, $r^2 = .76$); 2) all but the first 40 eigenvectors of the autoassociative memory ($r = .25$, $r^2 = .06$) 18 percent of the variance of the image is left unexplained [$1 - (.76 + .06) = .18$]. The bottom panels represent a spatial filtering analysis of the same face (low frequencies—preserving 94% of the power spectrum—*vs.* high frequencies—preserving 6% of the power spectrum). Filtering a *new* face through the autoassociative memory is quite different from filtering it through spatial frequency filters. The autoassociative memory filters the face image through the statistical properties of the set of learned faces and so distorts the new image in proportion to its difference from the set of learned faces.

to simulate gender categorization is detailed in the following section. For now, we simply note that the results of this preliminary simulation indicated that compressing the faces by local averaging with a 5×5 window caused relatively little change in the performance of the model on gender classification. This is made clear in Figure 5, which shows the proportion of correct classifications obtained with complete

FIGURE 3. The left panels show the original face. The center panels show: top panel – the face reconstructed with the first 40 eigenvectors of an autoassociative memory trained with 160 Caucasian faces. The correlation between the original face and the reconstructed face is: $r = .94$ (the explained variance is: $r^2 = .88$); bottom panel — the face after lowpass filtering (*i.e.,* only low frequencies are preserved), 93% of the image power is preserved. The right panels show: top panel — the face reconstructed with all but the first 40 eigenvectors of the autoassociative memory. The correlation between the original face and the reconstructed face is: $r = .36$ (the explained variance is: $r^2 = .13$); bottom panel – the face after high-pass filtering (*i.e.,* only high frequencies are preserved), 7% of the image power is preserved. Filtering a learned face through the autoassociative memory is somewhat similar to filtering it through spatial frequency filters.

shown (Harmon, 1973; Samal, 1991) that there is enough information in a 32×32 pixel image of face digitized with a resolution of 8 gray levels, for faces to be correctly identified by human subjects. To verify that we did not lose information essential for gender categorization, a preliminary set of simulations was performed for a sample of conditions using both complete and compressed faces. The method used

recognition and identification tasks require finer information than a simple categorization task, performance for these tasks can be improved by the addition of a high frequency range. High frequencies carry information concerning the inner features of a face (*e.g.,* specific shape of the eyes, nose, and mouth). Figure 3 illustrates the similarity between the principal component and the spatial frequency analysis for faces learned by the memory: Eigenvectors with large eigenvalues contain mostly low frequency information, and eigenvectors with small eigenvalues contain essentially high frequency information.

However, as noted by Abdi *et al.* (1995), the apparent similarity between these two approaches should not be interpreted as an indication that the PCA approach could be reduced to a simple spatial filtering technique. In contrast to spatial filters, eigenvectors depend on the statistical structure of the set of faces from which they are extracted. Figure 4, for example, shows that a Japanese face filtered through Caucasian eigenvectors is dramatically distorted, whereas spatial filtering does not have the same effect: When high frequencies are retained the face is recognizable. This property of eigenvectors has been used by O'Toole, Deffenbacher, Abdi & Bartlett (1991) to model the often cited other-race effect as a perceptual learning problem. They trained two autoassociative memories to reconstruct a large number of faces of one race ("own race") and a smaller number of faces of another race ("other race"). The ability of the memories to reconstruct new faces from both races was then tested. Results showed that new faces from the majority race were better reconstructed than new faces from the minority race. Moreover, reconstructions of new faces from the minority race were more similar to each other than reconstructions of new faces from the majority race.

In summary, the principal component and the spatial frequency approaches provide somewhat similar and complementary analysis of the information contained in facial patterns. The work by O'Toole *et al.* (1993), Valentin and Abdi (1996) and Sergent (1986a) shows that most of the information in faces as measured by traditional image analysis techniques (*i.e.,* energy spectrum for the spatial frequencies analysis, and variance explained or inertia for the principal components analysis) is conveyed by the lower frequency bandwidth or by the eigenvectors with the largest eigenvalues. This does not imply, although it was suggested by Ginsburg (1978) and Sirovich and Kirby (1987), that the information conveyed by the high frequencies or by the eigenvectors with relatively small eigenvalues is redundant or not useful. In fact, both the principal component and the spatial frequency approaches suggest the existence of a dissociation between two kinds of information in faces: 1) general configural information (basic shape and structure of the face) which is conveyed by low spatial frequencies or eigenvectors with large eigenvalues, and is useful for general semantic categorization; and 2) highly detailed, identity-specific information, which is conveyed by high spatial frequencies or eigenvectors with small eigenvalues, and is useful for face recognition and identification.

3. Simulations

In the first two series of simulations we tested the "generalizability" of the information captured by eigenvectors for classifying new faces (*i.e.,* not learned) according to their gender. The first series evaluates the ability of the model to generalize to new faces from the same race as the learned faces, and the second one the ability to generalize to new faces from a different race. The third series of simulations estimates the stability of the information carried by different eigenvectors as a function of their eigenvalues. These simulations were designed to extend previous analyses of the statistical properties of the eigenvectors extracted from the cross-product matrix of a set of male and female face images.

3.1. **Stimuli.** A set of 320 full-face pictures of young adults (80 Caucasian females, 80 Caucasian males, 80 Japanese females, and 80 Japanese males) was used as the database. The images were roughly aligned along the axis of the eyes so that the eyes of all faces were about the same height. None of the pictured faces had major distinguishing characteristics, such as beards or glasses. Each face was digitized from a slide as a $225 \times 151 = 33975$ pixel image with a resolution of 16 gray levels per pixel. For computational convenience, faces were compressed by local averaging, using a 5×5 window, giving $46 \times 31 = 1426$ pixel images.

Reducing the number of pixels in an image filters out part of the high detailed information, preserving only the low spatial frequency information. This should not be a problem, however, since it has been

FIGURE 2. The *top panels* show three original faces and the *bottom panels* the "responses" produced by an autoassociative memory trained with 160 Caucasian faces when these faces are presented as input after complete Widrow-Hoff learning. The faces are 1) a Caucasian face learned by the memory, 2) a Caucasian face that has not been learned by the memory, and 3) a Japanese face that has not been learned by the memory. The learned face is reconstructed perfectly by the memory ($r = 1$). The Japanese face is more distorted by the memory than the new Caucasian face ($r = .93$, $r^2 = .87$ versus $r = .82$, $r^2 = .68$ respectively).

When displayed visually (*i.e.,* as an image), the eigenvectors of the weight matrix appear face-like. They can be thought of as a set of "global features", "macrofeatures," or "eigenfeatures" from which the faces are built (Abdi, 1988; O'Toole & Abdi, 1989; Sirovich & Kirby, 1987). The estimation of a face by the system can, thus, be represented as a weighted sum of eigenvectors (from Eq. 2, and Eq. 5)

$$\widehat{\mathbf{x}}_k = \sum_{\ell=1}^{L} \lambda_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T \mathbf{x}_k = \sum_{\ell=1}^{L} \lambda_\ell \gamma_\ell \mathbf{u}_\ell \quad \text{with } \gamma_\ell = \mathbf{u}_\ell^T \mathbf{x}_k \tag{7}$$

where the weights γ_ℓ are the projections of the faces onto the eigenvectors. These weights can be interpreted as an indication of the extent to which a given eigenvector (or "macrofeature") characterizes a particular face.

When Widrow-Hoff learning is used, Eq. 5 reduces to

$$\mathbf{W} = \mathbf{U}\mathbf{U}^T \tag{8}$$

and the estimation of a face is obtained by dropping the eigenvalues in Eq. 7:

$$\widehat{\mathbf{x}}_k = \sum_{\ell=1}^{L} \gamma_\ell \mathbf{u}_\ell \quad \text{with } \gamma_\ell = \mathbf{u}_\ell^T \mathbf{x}_k \ . \tag{9}$$

Intuitively, this is equivalent to giving the same importance to each eigenvector in the reconstruction of a face. More formally, we say that Widrow-Hoff learning amounts to sphericizing the weight matrix $\mathbf{W}$.

A first advantage of using eigenvectors to represent faces is that they are determined *a posteriori.* A second advantage is that they reflect the statistical structure of the set of faces from which they are extracted. As an illustration, Figure 1 displays the first three eigenvectors extracted from a set of 80 male faces (left panels) and the first three eigenvectors extracted from a set of 80 female faces (right panels). Clearly, these two sets of eigenvectors differ in global shape and form.

From a signal processing point of view, the system acts somewhat like a Wiener filter (Abdi, 1994a). In neural network terminology this type of system is known as a linear content addressable memory. When new faces are presented as memory keys, they are filtered through the features extracted from the set of learned faces. Hence, new faces that resemble the learned faces are less distorted by the memory than new faces which are very different from the learned faces. In other words, the more different a new face is from the learned faces, the poorer the quality of reconstruction of this face will be. This is illustrated by Figure 2 (see also Valentin, Abdi, Edelman, & Nijdam, 1996, for a detailled analysis of this phenomenon).

2.3. **Analysis of Perceptual Information in Faces.** An interesting aspect of the PCA approach is that it provides a tool for analyzing the perceptual information in faces. For example, O'Toole *et al.* (1993), and Valentin and Abdi (1996) examined the kind of information provided by different ranges of eigenvectors, and the usefulness of this information with respect to specific 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. More specifically, they showed that the eigenvectors with larger eigenvalues convey information relative to the basic shape and structure of the faces as well as their orientation (*e.g.,* full-face or profile). This kind of information is most useful in categorizing faces along general semantic dimensions such as gender or race (O'Toole, Abdi, Deffenbacher & Bartlett, 1991). In contrast, the eigenvectors with smaller eigenvalues capture information that is specific to single or small subsets of learned faces. These eigenvectors are most useful for distinguishing a particular face from any other face.

These results can be related to some earlier work by Sergent (1986a, 1986b), in which facial information was analyzed in terms of spatial frequencies. Sergent reported that different kinds of facial information are conveyed by different physical characteristics of the face. Specifically, faces contain both featural and configural properties. These properties are not conveyed by the same spatial frequency ranges and are not equally useful, depending on the nature of the processes involved in performing a particular task. A low frequency representation provides information concerning facial configurations (*i.e.,* general shape, outer contour, and hairline of a face) but does not provide a detailed representation of the individual features. Consequently, categorization tasks may be accurately achieved by processing only the low frequency range of a face image. In contrast, because

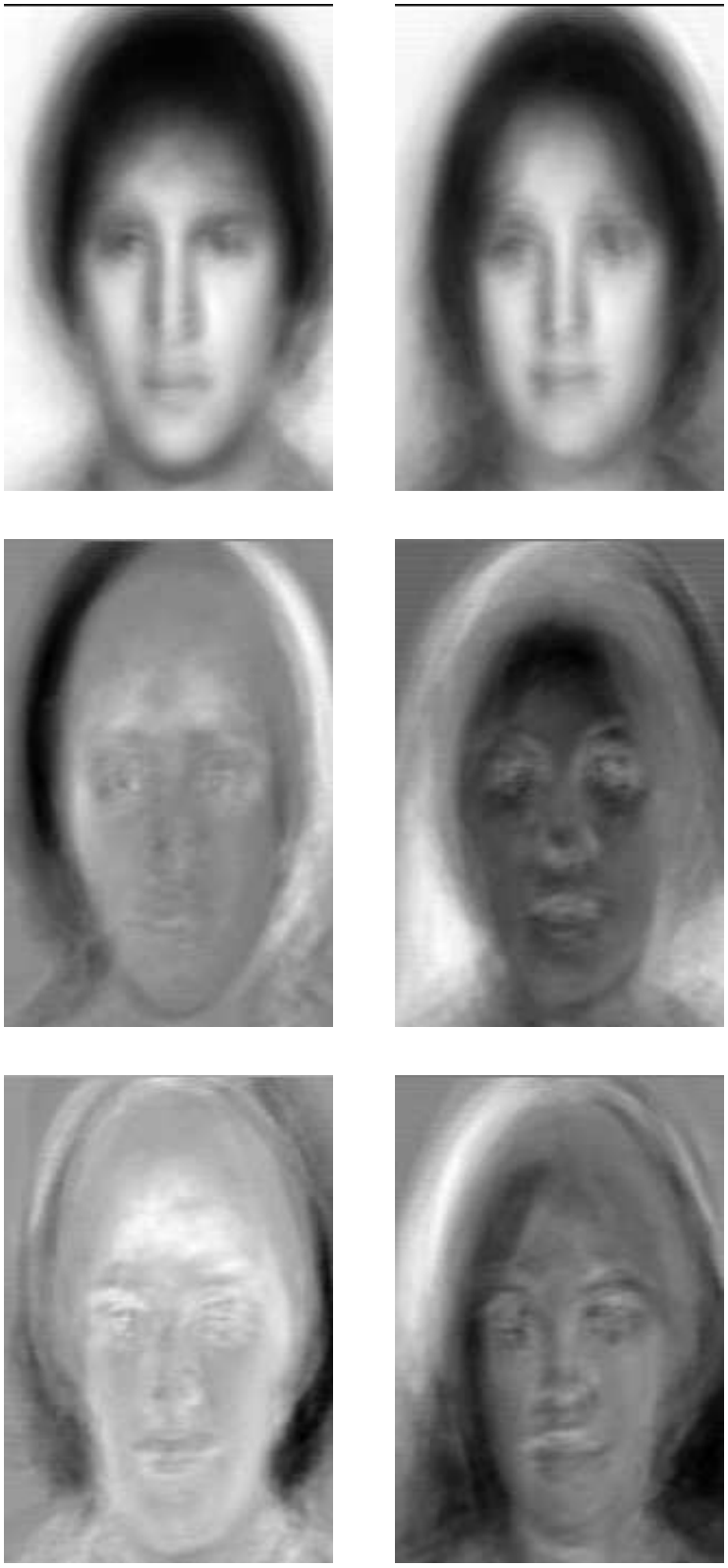

FIGURE 1. The first 3 eigenvectors of an autoassociative memory trained with 80 Caucasian male faces appear in the *left panels*, the first 3 eigenvectors of an autoassociative memory trained with 80 Caucasian female faces in the *right panels*. Global differences in shape and form can be observed between the two sets of eigenvectors.

memory is to find the values or weights for the 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. Kohonen (1977) used faces as stimuli to illustrate some properties of autoassociative memories. Specifically, he showed that an autoassociative memory can act as a content addressable memory for faces. Both Kohonen (1977) and Anderson *et al.* (1977) pointed out that using an autoassociative memory to store a set of patterns is equivalent to computing the eigen-decomposition of the cross-product matrix created from the set of features describing these patterns, or, in other words, performing the principal component analysis of the set of patterns. The model is presented first, followed by a discussion of the interpretation of eigenvectors as "macrofeatures".

2.1. **Model Description.** To construct the input patterns, each face is digitized and coded as a I-dimensional vector $\mathbf{x}_k$, concatenated from the columns of the face image (with I representing the number of pixels, and k indexing the faces). For example, if the kth face image is a 225×151 pixel image, it is represented by the 33975-element vector $\mathbf{x}_k$, where each entry represents a gray scale value. The vectors are normalized so that $\mathbf{x}_k^T \mathbf{x}_k = 1$. The set of K faces composing the learning set is represented by an $I \times K$ matrix $\mathbf{X}$ in which the kth column is equal to $\mathbf{x}_k$.

The faces are stored in an autoassociative memory composed of I neuron-like units as follows: Each unit is connected to all the other units, and the intensities of the connections are represented by an $I \times I$ matrix $\mathbf{W}$. When standard Hebbian learning is used, $\mathbf{W}$ is given by the sum of the outer product matrices of each face vector

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{x}_k \mathbf{x}_k^T = \mathbf{X}\mathbf{X}^T \ . \tag{1}$$

The reconstruction of the kth face is obtained by premultiplication of the vector $\mathbf{x}_k$ by the matrix $\mathbf{W}$:

$$\widehat{\mathbf{x}}_k = \mathbf{W}\mathbf{x}_k \tag{2}$$

where $\widehat{\mathbf{x}}_k$ represents the estimation of the kth face by the memory. The quality of this estimation can be measured by computing the cosine of the angle between $\widehat{\mathbf{x}}_k$ and $\mathbf{x}_k$:

$$\cos(\widehat{\mathbf{x}}_k, \mathbf{x}_k) = \frac{\widehat{\mathbf{x}}_k^T \mathbf{x}_k}{||\widehat{\mathbf{x}}_k|| \ ||\mathbf{x}_k||} \ . \tag{3}$$

A cosine of 1 indicates a perfect reconstruction of the stimulus.

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 expected response by iteratively changing the weights in matrix $\mathbf{W}$ as follows:

$$\mathbf{W}_{[t+1]} = \mathbf{W}_{[t]} + \eta(\mathbf{X} - \mathbf{W}_{[t]}\mathbf{X})\mathbf{X}^T \tag{4}$$

where η is a small positive constant (typically smaller than one).

2.2. **Eigenvectors as "Macrofeatures".** As pointed out by Anderson *et al.* (1977) and Kohonen (1977), since the weight matrix $\mathbf{W}$ is a cross-product matrix, it is positive semi-definite (*i.e.,* all its eigenvalues are positive or zero, and all of its eigenvectors are real). As a consequence, $\mathbf{W}$ can be expressed in a convenient way as a weighted sum of its eigenvectors:

$$\mathbf{W} = \sum_{\ell=1}^{L} \lambda_\ell \mathbf{u}_\ell \mathbf{u}_\ell^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \qquad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{5}$$

where $\mathbf{u}_\ell$ is the ℓ-th eigenvector of $\mathbf{W}$, λ_ℓ the ℓ-th eigenvalue, $\mathbf{I}$ stands for the identity matrix, $\mathbf{\Lambda}$ represents the $L \times L$ diagonal matrix of eigenvalues, $\mathbf{U}$ is the $I \times L$ matrix of eigenvectors, 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.

The eigenvectors and eigenvalues of the weight matrix $\mathbf{W}$ can be obtained directly using the singular value decomposition (cf., *e.g.,* Horn & Johnson, 1985) of the face matrix $\mathbf{X}$. Formally:

$$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \tag{6}$$

where $\mathbf{U}$ represents the matrix of eigenvectors of $\mathbf{X}\mathbf{X}^T$, $\mathbf{V}$ represents the matrix of eigenvectors of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{\Delta}$ is the diagonal matrix of singular values, which are equal to the square roots of the eigenvalues of $\mathbf{X}\mathbf{X}^T$ and $\mathbf{X}^T\mathbf{X}$ (they are the same).

of a set of learned faces. These features or "macrofeatures" (Anderson & Mozer, 1981) are the eigenvectors, or principal components of the pixel cross-product matrix of a set of faces. They can be obtained directly (Sirovich & Kirby, 1987; Turk & Pentland, 1991) or via a neural network–linear autoassociator (Abdi, 1988; Diamantaras & Kung, 1996) or a backpropagation network (Cottrell & Fleming, 1990). This approach, generally called the *principal component analysis* (PCA) approach has the advantage of eliminating the difficult problem of feature selection and extraction. Although not intended as a general solution to the problem of face processing, the PCA approach provides a way of modeling a wide range of tasks including face categorization and recognition, as well as simulating some well-known psychological phenomena such as the other-race effect[1] or the effect of typicality[2] on face recognition (cf. O'Toole, Abdi, Deffenbacher, & Valentin, 1995, for a review). Originally applied to raw pixel-based representation of faces, this approach has been applied recently to more sophisticated representations such as 2D separated shape and texture representations (*e.g.,* Hancock, Burton & Bruce, 1996) and 3D laser scan data (*e.g.,* O'Toole, Vetter, Troje, & Bülthoff, 1997).

The general purpose of the present work is to analyze the robustness of the kind of face representation proposed by the PCA approach. The human face is a complex visual pattern that contains general categorical information as well as idiosyncratic, identity specific information. By categorical information, we mean that some aspects of a face are not specific to that particular face but are shared by subsets of faces (*e.g.,* female faces share some visual characteristics such as smoothness of the skin, prominence of the cheeks, or roundness of the face). These aspects can be used to assign both unfamiliar and familiar faces to general semantic categories such as gender or race.

Previous work showed that the PCA approach dissociates spontaneously between different types of information. O'Toole, Abdi, Deffenbacher and Valentin (1993) reported that eigenvectors with large eigenvalues capture information that is common to subsets of faces (*i.e.,* categorical information) and eigenvectors with small eigenvalues capture information specific to individual faces (*i.e.,* identity specific information). In a recent study, Abdi *et al.* (1995) showed that, while all eigenvectors are necessary to categorize learned faces optimally along general categories (*e.g.,* gender), only the information contained by the eigenvectors with large eigenvalues can be generalized to *new* faces. Our objective was to examine further the generalizability and the stability of categorical information conveyed by the eigenvectors of a set of male and female faces.

The present paper is organized as follows. The PCA approach is presented briefly first (for a more detailed presentation see *e.g.,* Valentin, Abdi, O'Toole & Cottrell, 1994 or Valentin, Abdi & O'Toole, 1994) followed by a discussion of the interpretation of eigenvectors as "macrofeatures." Next, the usefulness of this approach for analyzing the perceptual information in faces is discussed along with its relationship to some earlier work on the role of different spatial frequencies in face processing. Finally, three series of simulations concerning the statistical properties of eigenvectors derived from a set of face images are described. In the first series, we estimate the generalizability of the gender information conveyed by the eigenvectors extracted from small sets of faces. In the second series, we examine if the gender information extracted from a given population of faces (*e.g.,* Caucasian) can be generalized to faces from a different population (*e.g.,* Japanese). In the third series, we evaluate the "stability" of face eigenvectors (*i.e.,* the minimum number of faces necessary to estimate them) as a function of the variance they explain in the set of faces. The results of these simulations are then discussed in relation to temporal properties of the visual system and are put in perspective with some neuropsychological data.

2. Principal Component and Linear Neural Network Approach

This approach is based on earlier work by Anderson, Silverstein, Ritz, and Jones (1977) and Kohonen (1977) on autoassociative memories. Autoassociative memories are a special case of associative memories in which the input patterns are associated with themselves. The goal of constructing an autoassociative

[1] the fact that it is easier to recognize faces from our own race than from another race.

[2] the fact that typical (*i.e.,* average) faces are harder to recognize than distinctive faces.

Vol.42 , No.4, December 1997, pp. ***–***. / Journal of Mathematical Psychology

Principal Component and Neural Network Analyses of Face Images: What Can Be Generalized in Gender Classification?

Dominique Valentin[*†], **Hervé Abdi**[*†], **Betty Edelman**[*] **and Alice J. O'Toole**[*]
[*] *The University of Texas at Dallas,* [†] *Université de Bourgogne à Dijon*

We present an overview of the major findings of the principal component analysis (PCA) approach to facial analysis. In a neural network or connectionist framework this approach is known as the linear autoassociator approach. Faces are represented as a weighted sum of macrofeatures (eigenvectors or eigenfaces) extracted from a cross-product matrix of face images. Using gender categorization as an illustration, we analyze the robustness of this type of facial representation. We show that eigenvectors representing general categorical information can be estimated using a very small set of faces and that the information they convey is generalizable to new faces of the same population and to a lesser extent to new faces of a different population.

1. Introduction

One of the major problems in modeling face processing is to find a way of representing faces that allows for the wide range of tasks typical of human performance. Traditionally, computational models of face recognition represent faces in terms of geometric descriptors that include distances, angles, and areas between elementary features such as eyes, nose, or chin (Harmon & Hunt, 1977; Harmon, Khan, Lash & Ramig, 1981; Kaya & Kobayashi, 1972; Sakai, Nagao & Kidode, 1971) or in terms of template parameters (Yuille, 1991), or isodensity lines (Nakamura, Mathur & Minami, 1991). Although these approaches economically represent faces in a way that is relatively insensitive to variations in scale, tilt, or rotation of the faces, they are not without problems (for a review, see Samal & Iyengar, 1992).

The major difficulty with representing faces as a set of features is that it assumes some *a priori* knowledge about what are the features and/or what are the relationships between them that are essential to the task at hand. Burton, Bruce, and Dench (1993), for example, showed the difficulty of finding a set of features useful in discriminating accurately between male and female faces. In a series of five experiments, they examined the usefulness of different kinds of feature measures for predicting the gender of a set of faces. The measures they used ranged from simple raw distances between facial landmarks (*e.g.,* pupils) to more complex measures including ratios of distances, or angles taken from full faces and/or profile views of the faces. They showed that no simple set of features can predict the gender of faces accurately. By combining measurements from all five of their experiments, however, they obtained 94% correct classification by gender (for the learning set). They concluded that "explicit measurement" of facial features is probably not the best "basis for automated face recognition systems".

Abdi, Valentin, Edelman, and O'Toole (1995), showed that comparable gender categorization performance can be obtained using *a posteriori* features automatically derived from the statistical structure

Thanks are due to June Chance and Al Goldstein for providing the faces used in the simulations, and to Mike Burton for helpful comments on an earlier draft of this paper. Alice O'Toole is supported by NIHM grant 1R29MH5176501A1.
Correspondence about this paper should be send to Dominique Valentin, ENSBANA, 1, Place Erasme, Dijon 21006 Dijon Cedex, France; E-mail: *valentin@u-bourgogne.fr* or to Hervé Abdi, Program in Applied Cognition and Neuroscience, The University of Texas at Dallas, Richardson, TX 75083-0688, U.S.A.; E-mail: *herve@utdallas.edu.*