Perception, 1995, volume 24, pages 539–562

More about the difference between men and women: evidence from linear neural networks and the principal-component approach

Hervé Abdi†, Dominique Valentin, Betty Edelman, Alice J O'Toole

School of Human Development, The University of Texas at Dallas, Richardson, TX75083–0688, U.S.A.,
†also Université de Bourgogne, 21004, Dijon, France.
Received 20 December 1994.

Abstract. The ability of a statistical/neural network to classify faces by sex by means of a pixel-based representation has not been fully investigated. Simulations using pixel-based codes have reported sex classification results that are less impressive than those reported for measurement-based codes. In no case, however, have the reported pixel-based simulations been optimized for the task of classifying faces by sex. A serie of simulations is described in which four network models were applied to the same pixel-based face code. These simulations involved either a radial basis function network or a perceptron as a classifier, preceded or not by a preprocessing step of eigendecomposition. It is shown that performance comparable to that of the measurement-based models can be achieved with pixel-based input (90%) when the data are preprocessed. The effect of the eigendecomposition preprocessing of the faces is then compared with spatial frequency analysis of face images and analyzed in terms of the perceptual information it captures. It is shown that such an examination may offer insight into the facial aspects important to the sex classification process. Finally, the contribution of hair information to the performance of the model is evaluated. It is shown that, although the hair contributes to the sex classification process it is not the only important contributor.

1. Introduction

Human beings are able to make accurate and fast visually-based categorizations. A particularly striking example of this ability comes from studying face processing and categorization. Among the variety of face categorization tasks, sex classification is one of the most biologically important and probably the easiest and fastest to achieve. Bruce et al (1987) found that, on the average, only 600 ms were necessary for classification of faces according to their sex. In a more recent study, Bruce et al (1993) reported that human observers were able to classify photographs of non familiar faces with respect to sex with 96% accuracy, even though the hair was concealed by a swimming cap.

However, the apparent ease with which human beings are able to perform this task should not mask its complexity. As pointed out by Bruce et al (1993), the exact nature of the visual information used by human beings to decide whether a face is male or female is far from being understood. Recent attempts at creating automatic face-recognition devices have made this point clear. A first and major issue in these computer models concerns the choice of how to code the input faces. One strategy for automatically classifying faces by sex consists of selecting a priori a set of descriptors

and testing whether these descriptors can be used to predict the sex of faces. Even though this approach has yielded models that perform reasonably well on the task of gender classification (cf Brunelli and Poggio 1992; Burton et al 1993) it is not completely satisfactory. As noted by Burton et al, it requires the selection of a set of explicit measurements that must be located (either by hand or automatically) in the faces prior to using them as input to the automatic gender classifier. In addition, from a psychological point of view, these codings discard textural information like the "smoothness" of faces that may be used by humans in making a gender classification. In their discussion section, Burton et al suggested that a low level input code, like a pixel code, might provide a better way of representing faces.

The major reason given by Burton et al for not using a low level coding was the relatively poor gender-classification performance reported in the previous literature when this type of coding had been used. For example, authors such as Cottrell and Fleming (1990) and O'Toole et al (1991) have reported 63% and 74% correct gender classification, respectively, in their models. However, these models were trained under conditions that make a comparison of their results with those of the measurement-based models difficult. Specifically, the model of Cottrell and Fleming was trained to classify images simultaneously by "faceness", sex, and identity by using a very small number of faces, and O'Toole et al used only a small amount of the information available for the gender-classification task in their model. Therefore, realistically assessing the level of sex-classification performance that can be obtained by a statistical/neural model involving a low-level coding scheme remains open for investigation.

In this paper, after a brief overview of previous models, we present some new simulations designed to initiate a more systematic evaluation of the use of a low-level code. We begin by reporting a simulation in which we implemented several neural network algorithms operating on pixel-based codes for the task of gender classification. We do this with the purpose of examining the gender-classification potential of low level codes under more optimal conditions than have been tested previously (ie by using all available information and applying this to a dedicated sex-classification network). While we realize that comparisons among different gender-classification models are never entirely satisfactory owing to differences in the size, composition, and diversity of the stimulus set, we show here that pixel-based codes can support accurate gender classification.

The second purpose of this study was to evaluate the utility of subcomponents derived from the pixel-based code in the neural network (via the eigendecomposition). In previous work these subcomponents have been related to the recognizability and visually based categorical structure of faces (eg O'Toole et al 1994). The analysis we present here is focused on the contrast between components that carry gender information that generalizes to novel faces and components that the neural networks exploit for learned faces, but which do not generalize to novel faces.

2. Previous work

We begin with a brief overview of some previous work in the modeling of face processing, which can be considered to fall into two broad categories based upon the type of representation used: measurement-based or pixel-based coding of the faces. Summary information for the models that we discuss in this section is presented in Table 1.

2.1. *Measurement-based models.* Traditionally, statistical/computational models of face processing have represented faces in terms of distances, angles, and areas between elementary features such as eyes, nose, or chin (cf Samal and Iyengar 1992). A recent study by Brunelli and Poggio (1992) indicates that such a geometrical coding scheme can be used successfully to categorize faces by gender. They used a set of 16 geometrical features (eg pupil-to-eyebrow separation, nose width,

Authors	Architecture	Size of training set	Percentage correct	Size of testing set	Percentage correct
Models involving measurement-based coding if the faces					
Brunelli et al. 1991	Hyper basis function networks (19 features)	40	90	42	70
Burton et al. 1993	Discriminant analysis (16 features)	179	94	—	—
Models involving a pixel-based coding if the faces					
Cottrell et al 1991	Compression network (40hu)+classification net	64	98	78	63
Golomb et al 1991	Compression network (80hu) + classification net	80	100	10	91.9
O'Toole et al. 1991	Autoassociative memory (first 4 eigenvectors)	80	74	—	—

Note: "hu" stands for "hidden units."

TABLE 1. Principal statistical/neural network models used to categorize faces by sex.

and chin radii) as input to two competing hyper basis function (HBF) networks[1], one trained on male faces, the other one on female faces. The features were first extracted automatically from a data base of 168 face images obtained by taking four views of 42 faces (21 males and 21 females). Each network was composed of 16 input units connected to a "center" (hidden unit) converging to 1 output unit. During learning, the networks developed a "prototype" of the faces on which they were trained. The classification task was simulated by presenting a face vector as input to both networks. The face was classified according to the HBF network with the strongest activation. Thus, if the response of the male network was greater than the response of the female network, the face was categorized as male, otherwise it was categorized as female.

The performance of the HBF networks was evaluated both for old and for new views of the faces from the training set, and for new faces where no view was learned before. The ability of the network to generalize to new views of old faces or new faces was examined. Sex categorization of faces was tested using each view, or possibly all four views for new faces (Brunelli and Poggio did not mention explicitly the number of views used to test new faces), in turn. Results showed that 90% of the old views of old faces, 86% of new views of old faces, and 79% of the new faces were classified correctly. As a comparison, the performance of human subjects on the same set of faces was 90% correct[2].

Brunelli and Poggio's results suggest that information useful for categorizing faces according to their sex can be captured by a very limited number of geometrical measurements. The major difficulty, however, with this type of approach is that there is no *simple* algorithm for selecting and extracting the features essential for the task at hand (Brunelli and Poggio 1993).

Burton et al (1993) illustrate the difficulty of finding a reliable set of features useful for discriminating accurately between male and female faces. In a series of five experiments, they examined the usefulness of different kinds of facial measurements for predicting the sex of a set of faces. In their

[1]The hyper basis function network is a generalization of the radial basis function network presented in appendix 3.

[2]To produce a response from the human subjects that was solely visually derived, all the faces were, obviously, new to them.

first experiment, Burton et al used an ascendant stepwise discriminant analysis to predict the sex of a set of 179 faces (91 male and 88 female faces) using 19 variables derived from 73 measurements taken on frontal views of the faces. The 19 variables were chosen because they seemed relevant both to experts designing data-retrieval systems and to eyewitnesses describing faces. These variables represented either between-features distances (eg distance between eyebrows, eye-to-eyebrow distance) or feature measurements (eg mouth width, forehead height). The variables were all standardized for head size by division by the interocular distance. Keeping 12 of the original variables, they achieved a highly significant level of correct categorization of 85.7% for the male faces and 85.2 % for the female faces. A more fine-grained analysis showed that the variables mostly responsible for this level of correct classification were eyebrow thickness, width of the nose, eye-to-eyebrow distance, forehead height, and distance between corners of the eyebrows.

In additional experiments, Burton et al used a similar methodology with more sophisticated variables [ie what Rhodes (1988) refers to as second-order or configural variables] such as angles (eg point of chin from jaw at mouth line) and ratio of distance (eg mouth width to distance between mouth and nose) constructed from the same set of 73 measurements. Even though we might expect a better performance with these kinds of variables, curiously their use in this case did not lead to a better prediction of gender (69.2% correct categorization for the male faces and 77.3% for the female faces), nor did a set of second-order variables derived from profile views of the same faces (64.8% correct categorization for the male faces and 64.8 % for the female faces). Using three-dimensional distances, on the other hand, gave a level of performance comparable to the success rate of variables of the first experiment derived from the frontal view (82.4% correct categorization for the male faces and 88.6% for the female faces). Combining all these measurements in a final stepwise ascendant discriminant analysis gave the best performance of all (93.4% correct gender classification for the male faces and 94.3% for the female faces).

In summary, the Burton et al (1993) study showed that, although it is possible to find a linear combination of measurements that can be used to discriminate reliably between male and female faces (about 90% accuracy), this approach is not without problems. First, the authors demonstrated that no simple combination of features was able to predict the sex of faces with a performance level comparable to human subjects. Second, they reported that the variables useful for the discriminant analysis did not correlate highly with gender distinctiveness as rated by human observers, nor did the human observers and discriminant analysis agree on misclassifications or criteria for classifying faces by gender (even when their level of performance was very close). This last point may indicate that human subjects use some kind of facial information that is not preserved in a feature-based representation of the faces (eg textural information). Actually, given the intensive effort required to preselect and code features, and some lack of agreement with human data, Burton et al suggested that explicit measurements of facial information may not be the best "basis for automated face-recognition systems" (1993, page 174).

2.2. *Models based on principal-component analysis.* As noted by Burton et al, an alternative approach to explicit measures, which has been used in several computational models of face processing, is to operate directly on pixel-based codings of the faces. When codings of this sort have been used, they have generally been implemented in conjunction with a model that computes (via a linear autoassociator) or approximates (via a back-propagation compression network) the eigendecomposition of the pixel cross-product matrix of the set of faces on which it is trained. The eigenvectors, (or the representation extracted by the hidden layers of a compression network), are then used as input to a classification network. This approach is often referred to as the principal-component-analysis or PCA approach.[3]

[3]For a detailed review of this approach to face processing, cf. Abdi and Valentin 1994; Valentin et al. 1994a, 1994b.

Cottrell and Fleming (1990) used a two-stage network to classify a set of face images according to "faceness", sex, and identity. A training set composed of 64 face images (5 or 6 different views of 11 faces) and 13 nonface images were first compressed via a 3-layer back-propagation autoassociative network (ie one input layer, one hidden layer, and one output layer). The representation formed in the hidden units was then used as input to a two-layer network trained to classify the face images. When tested with previously learned images, the model attained perfect performance on the assignment of "faceness", sex, and identity. New images of faces were categorized perfectly with respect to "faceness", but not according to sex (37% error: 26 out of 70 female images were classified as male, whereas all the male faces were correctly classified). This failure to generalize to new faces, however, was most probably due to the small number of different faces used in the training set (only 11 faces) rather than to the type of face representation used. As a confirmation of this conjecture, Golomb et al (1991), using a similar approach, with a larger learning set (80 faces) and a classification network dedicated to sex categorization (SEXNET), obtained better gender-classification performance for new faces (91.9% correct classification averaged across trials). Interestingly, the authors indicate that a control back-propagation-classification network, trained directly on the pixel images (ie without using the compression step), was able to categorize perfectly the old faces, but was unable to generalize to new faces[4]. This latter result suggests that the compression step is an important factor in the performance of SEXNET. However, in a subsequent study, with the same faces, Gray et al (in press) report that a perceptron trained directly on the face images achieved a performance of 78.9% on the new faces. Therefore, additional information is necessary to evaluate more thoroughly the effect of the compression step.

The work of Golomb et al (1991) shows that, when used with a training set of a reasonable size, the two-stage network originally proposed by Cottrell and Fleming (1990) is able to classify faces by sex with an accuracy similar to that of the measurement-based models. However, one major problem with this approach is that back-propagation compression networks are extremely computationally intensive (requiring, for example, about 2000 iterations to learn a face, in the case of Golomb et al 1991). In addition, an analysis of the network activity during learning showed that when a low learning rate is used, the hidden-unit outputs tend to stay within the linear range of the logistic function (Cottrell and Metcalfe 1991). This suggests that the face compression is basically a linear problem, thus making the nonlinear function of the hidden units unnecessary. Therefore, it might be more practical to achieve a linear compression of the faces (ie eigendecomposition of the set of faces), either directly or by using a linear autoassociator, and use the eigenvectors as input to a classification network.

Other studies, (Abdi 1988; O'Toole et al 1991; 1993) have shown that some of the eigenvectors (essentially the ones with large eigenvalues) of an autoassociative memory matrix, made from a heterogeneous set of faces, capture information that is useful for predicting the sex of the faces. For example, O'Toole et al (1991) reported that a combination of the first four eigenvectors extracted from a memory matrix, made of Caucasian and Japanese male and female faces, yielded correct sex predictions for 74.3% of the faces. Despite the small magnitude of this proportion compared with that for human observers and measurement-based model performance, these results are promising since (i) only a small portion of the relevant information was used [ie four eigenvectors, which is equivalent to using only the information extracted by a small number of hidden units in the Cottrell and Fleming (1990) and Golomb et al (1991) models]; and (ii) the training set was composed of faces from two different races. Having a set of faces composed of different races makes sex identification more difficult because it increases the variability of the set of faces.

In summary, the previous work on sex categorization of faces involving low-level representation seems promising and clearly merits further investigation. A more systematic evaluation of this

[4]The authors state this point as an observation, and do not report quantitative results.

technique will provide a better basis for comparing the results obtained with those produced by measurement-based models and human data.

3. Model-assessment criteria

Before we describe our simulation, it is important to consider two important factors for evaluating the performance of a model. First, the ability of the model to generalize its gender-classification performance to faces that were not part of the training set must be considered. Second, it is important to have a baseline for what might be considered "good performance". We discuss each of this issues in turn.

The ability of a model to generalize its categorical learning to stimuli not included in the training set is an important measure of its value. This quality of prediction for new observations can be estimated by several methods. One approach is to compute confidence limits, on the basis of the performance attained with the learning set, or to use other statistical techniques to provide an approximation of the quality of the prediction (eg stepwise regression). However, these techniques are quite sensitive to certain statistical assumptions (eg homoscedasticity, random sampling) which are difficult to meet for most training samples. In addition, when dealing with a large set of predictors compared with the number of observations, the problem of multicolinearity occurs. This refers to the fact that having more descriptors than objects to describe makes it possible to have a perfect prediction on the sample, but with no guarantee of how the prediction will do with new observations (this is also referred to, in the neural-network literature, as overfitting).

Another way of testing the ability of a model to generalize is to use samples of new observations to derive a prediction estimate. This can be achieved without having to increase the number of stimuli by using a leave-one-out jackknife technique (Efron 1979). In brief, if one has a data set of N faces, training can be performed on $N-1$ faces, leaving one face out for testing. After completion of learning, the sex of the new face is predicted. This procedure is executed N times. Each time, a different face is omitted from the learning set and subsequently used as a new face.

In addition, if the performance of different models is to be evaluated, a baseline is useful as a point of reference. A first possible baseline would be a random performance (ie 50%). However, because all the classifiers considered so far use supervised learning (ie they use the knowledge of the sex of the faces in the learning set to optimize performance) a better baseline would be the performance of an unsupervised algorithm (ie an algorithm that tries to cluster faces together solely on the basis of their physical similarities, without using the knowledge of their sex). Such an algorithm would give an indication of the natural grouping of the faces in the set. The k-means algorithm (also known as ISODATA in pattern recognition, cf Nadler and Smith 1993) provides such a vehicle.

The goal of k-means is to partition optimally a set of observations into k classes. A brief description of the algorithm follows. The algorithm begins by setting the coordinates of k centers (or means) to random values (hence the name of k-means). Each observation in the set is then assigned to the class represented by the closest center. Then, the mean coordinates for each class are computed and used as new centers from which to reiterate the procedure. The process stops when the assignment of the observations remains stable over time. As a comparison reference for the supervised models, we carried out a series of 1000 k-means classifications with $k = 2$ classes. Gender classification was "estimated" by selecting one group as corresponding to its majority gender. Doing so gives an indication of how the faces tend to segregate "spontaneously" into males and females. The mean correct gender classification obtained with the k-means algorithm was 77% which, in itself, indicates a strong "natural" separation of faces by gender. Prediction was better for male faces than for female faces, however (87% *vs.* 66%, with a median of 91% and 66% respectively). This difference in performance for male and female faces can be explained by the fact that, in our sample, male faces are more similar to each other than are female faces. In fact, some female faces,

(the ones with short hair) are, in some ways, more similar to male faces than they are to female faces with long hair. These faces are the ones that tend to be misclassified by the k-means algorithm. We will come back to this problem of the effect of hairstyles on the model performance later on.

4. Simulation

This simulation was performed to evaluate the potential of pixel-based codes, under optimal conditions, to support gender classification. As noted by Burton et al (1993), two different aspects of a model can be responsible for its level of performance: the code used to represent the faces and the algorithm used to achieve the categorization task. In order to be able to evaluate the importance of the code, independently of the statistical/neural model used for categorization purposes, it is necessary to use the same entry code with different classification models. Therefore, we selected two classifiers, those previously used to assess the usefulness of measurement-based codings of faces. As a first classifier we used a radial basis function (RBF) network, which is a specific implementation of the hyper basis function network used by Brunelli and Poggio (1992; cf Michelli 1986). The second classifier used in this paper is a perceptron, which is formally equivalent to the discriminant analysis used by Burton et al (cf Abdi 1994a). A formal description of these two models is provided in Appendix 3 and 2, respectively.

In general, models involving a low-level code first represent faces as projections on statistically derived dimensions known as principal components or eigenvectors: this is equivalent to compressing or filtering the face images. To evaluate the utility of this type of face preprocessing, we tested the performance of each classification network with and without using such a preprocessing. When no preprocessing was used, the classifiers were trained directly on the pixel-code (ie the pixel values were used as input). When a preprocessing was used, the projections of the faces onto the eigenvectors were used as input to the classifiers. Both learned and new faces were projected onto the eigenvectors derived from the set of learned faces. In the neural network literature the preprocessing step is, in general, implemented via a linear autoassociative memory trained with Widrow-Hoff learning (cf Appendix 1).

As illustrated by figure 1, four different networks were used to categorize faces according to their gender:

- *model 1* was an autoassociator (ie eigendecomposition) followed by a RBF network that operates on the projections of the faces onto the eigenvectors of the memory matrix;
- *model 2* was an autoassociator (ie eigendecomposition) followed by a perceptron that operates on the projections of the faces onto the eigenvectors of the memory matrix;
- *model 3* was a RBF network trained directly on the pixel representation of the faces;
- *model 4* was a perceptron trained directly on the pixel representation of the faces.

For model 4, two types of learning were used: simple Hebbian learning (model 4a) and Widrow-Hoff error-correction learning (model 4b). Using the Widrow-Hoff learning rule with a perceptron is almost equivalent to equalizing the eigenvalues of the weight matrix (it is equivalent to computing the Moore-Penrose pseudo-inverse, see appendix 2). Hence, a perceptron, trained directly on the raw images with Widrow-Hoff learning, can be expected *a priori* to perform in the same range as a perceptron following preprocessing (model 2). Such a result would indicate that an important aspect of the preprocessing step resides in the *orthogonalization* of the input set.

4.1. *Stimuli.* 160 full-face pictures of young Caucasian adults (80 females and 80 males) were digitized from slides using 16 gray levels to create $151 \times 225 = 33,975$ pixel images. 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 a beard or glasses. For all simulations, each face was coded as a 33,975 pixel vector denoted $\mathbf{x}_k$ (where k stands for the

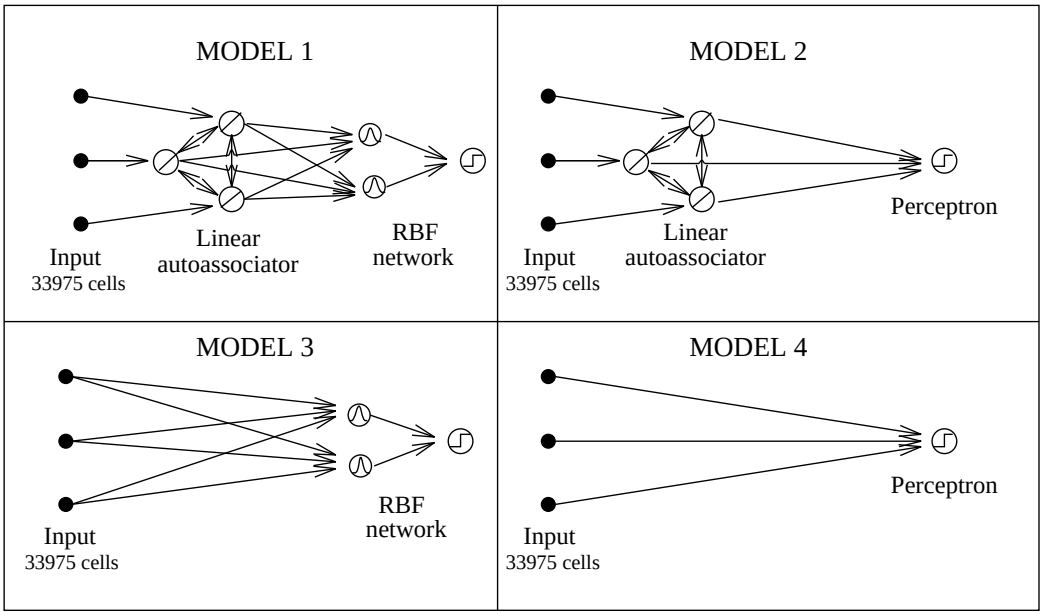

FIGURE 1. Four network models used to categorize faces according to their sex (RBF, radial basis function)

face). The vectors were obtained by concatenating the rows of the face images. For convenience, the vectors were normalized such that $\mathbf{x}_k^T \mathbf{x}_k = 1$ (for all k).

4.2. *Procedure.* In the following simulations, the jackknife technique previously explained was applied to our sample of 160 faces. This was done by iteratively training the networks (autoassociative memory and/or classifiers) with 159 faces, reserving one different face each time for testing purposes.

To implement the preprocessing stage in models 1 and 2, a $33,975$ unit autoassociative memory was created in which the 159 faces of the training set were stored. The weight matrix was then decomposed into its 159 eigenvectors (see appendix 1 for computational details). The projections on these eigenvectors of both the test face and the training faces were computed. Thus, each face was coded by a series of 159 projections corresponding to the "loadings" of the faces on the principal components of the weight matrix. The coefficients of the faces of the training set were then used to train the classification network (ie perceptron or RBF network). The prediction of the gender of the test face was made *after* learning was complete. Hence, the test face was not used to compute the eigenvectors, nor was it used to compute the optimum weights of the classification network.

For the classification stage in model 1, an RBF network was trained to associate the value +1 with female faces and the value of 0 with male faces. As a decision rule, faces associated with an output larger than .5 were classified as female, whereas faces associated with an output smaller than .5 were classified as male.

In model 2, the classification stage was implemented by a perceptron with complete Widrow-Hoff learning. In the specific case of two categories, this is equivalent to computing the barycenter (center of gravity) of each class and then computing the distance to both of these centers for the face to be classified. In terms of neural networks, this is done by training the perceptron to associate a value of 1 with female faces and a value of 0 with male faces. The decision rule is the same as the one used for the RBF network.

For models 3 and 4, the same classification networks were directly trained to classify the faces from the training set by using the $33,975$ dimensional pixel-vectors as their input. The prediction of the gender of the test face was made *after* learning was complete.

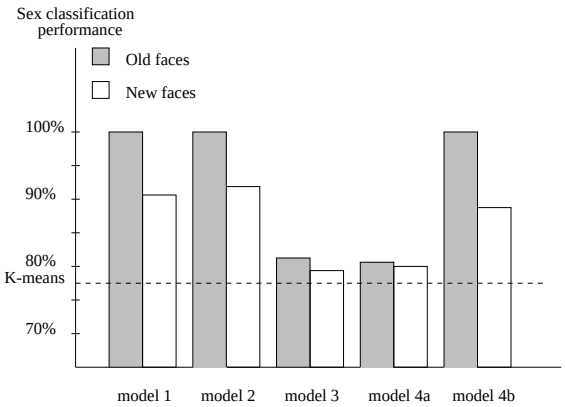

FIGURE 2. Percentage of correct sex classification achieved by the five network models for old and new faces. The models are as follows: 1, an autoassociative memory followed by a radial basis function network; 2, an autoassociative memory followed by a perceptron; 3, a radial basis function network trained directly on the pixel representation of the faces; 4a, a perceptron trained directly on the pixel representation of the faces by using a Hebbian learning rule; 4b, a perceptron trained directly on the pixel representation of the faces by using a Widrow-Hoff error-correction learning rule.

4.3. *Results.* The proportion of correct sex classification achieved by each network is presented in figure 2. As a baseline, the performance obtained using k-means is indicated with a dashed line. For the old faces, performance was averaged across the trials.

For the old faces, it can be seen from figure 2 that all models involving Widrow-Hoff learning (ie models 1, 2, and 4b) achieved 100% correct sex classification. This performance, however, should not be considered impressive: because, in this case, we are classifying 159 faces by means of $33,975$ descriptors, the descriptor set is obviously multicolinear, and hence, a model using an error-correction algorithm is guaranteed to reach 100% correct classification. In contrast, those models that did not include preprocessing or Widrow-Hoff learning in the classifier (models 3 and 4a) show performance only slightly better than that of the unsupervised k-means (81.1% and 80.9% respectively).

For the new faces, the best performance of 91.8% was achieved by a perceptron classifier preceded by the autoassociator eigendecomposition (model 2). An RBF using the same input (model 1) achieved an equivalent level of performance (90.6%). A perceptron using the Widrow-Hoff learning rule without any preprocessing (model 4b) attained 88.7% correct classification. As expected, this performance is close to that of model 2. Both the RBF without preprocessing (model 3) and the perceptron trained with Hebbian learning (model 4a) performed in the same range as the unsupervised performance baseline (79.3% and 80% respectively). Last, it is worth noting that for those latter models, the performance obtained for old and new faces is equivalent.

4.4. *Discussion.* The findings of the previous simulation add further support to the idea that pixel-based codings can yield accurate gender classification. Moreover, the performance superiority for

Models 1 and 2 suggests that preprocessing the face images via an eigendecomposition not only saves processing time by drastically reducing the size of the classifier networks, but also may find a set of relevant features. Further, the fact that a perceptron trained directly on the pixel-code by using Widrow-Hoff learning (model 4b) performed in the range of models 1 and 2 gives insight into the inner workings of the preprocessing step. As noted previously, using a Widrow-Hoff learning rule in a perceptron, or in an autoassociative memory, is equivalent to sphericizing the weight matrix (ie setting all eigenvalues equal to 1). This amounts to orthogonalizing the input set. Orthogonalization of the input signal is, incidently, a technique frequently used in signal processing in order to improve the performance of a system. An advantage of a separate preprocessing step to achieve this orthogonalization is that the output of this process is available for additional analysis, such as that described in the next section.

In addition, we showed that the particular classifier network used is not a critical factor in gender-classification performance. Specifically, the fact that a simple perceptron network was as efficient as a more complex RBF network suggests that gender discrimination is a linearly separable categorization task.

In conclusion, using a low level pixel representation of the faces in conjunction with an eigendecomposition preprocessing allows a simple neural network to achieve a level of performance superior to that of the unsupervised k-means. This level of performance compares favorably to that of Burton et al (1993). While these authors did not evaluate explicitly the performance of their model in generalizing to new faces, it is likely that the stepwise approach they followed yields, at least, a rough indication of the generalizability of their results (ideally, a good comparison between the approaches would be to use a jackknife methodology with Burton et al measurements).

5. A closer look at the preprocessing

Having illustrated the utility of the eigendecomposition preprocessing step in attaining the best gender-classification performance, we wished to take a closer look at this preprocessing step. Because computing the eigendecomposition of a set of faces is equivalent to performing the PCA of the pixel values (ie a Q-principal component analysis), standard statistical analysis techniques can be used to gain understanding of the role of the eigenvectors. Using this statistical framework, Sirovich and Kirby (1987) showed that (i) faces can be reconstructed to very recognizable levels using a subset of the eigenvectors—specifically, those with the largest eigenvalues; (ii) when visually displayed, the eigenvectors of a face matrix span the entire face and appear face-like. In a way, the eigenvectors can be interpreted as macro-features (cf Anderson and Mozer 1981) in the sense that each face can be rebuilt by a (linear) combination of the eigenvectors (see appendix 1).

In a subsequent study, O'Toole et al (1993) showed that, although eigenvectors with large eigenvalues were better sex predictors than those with small eigenvalues, a low dimensional representation of faces using eigenvectors with small eigenvalues, actually provided more distinctive identity-specific-information. As an illustration, figure 3 displays from left to right: 1) a female face, 2) its reconstruction with the first 20 eigenvectors, and 3) its reconstruction with eigenvectors 20 to 100. It is interesting to note that, although the first reconstruction explains most of the variance in the face ($r^2 = 0.81$), most human observers would have a problem identifying it, but would be able to recognize its sex. The second reconstruction, even though its correlation with the original face is quite low ($r^2 = 0.17$) can be easily identified in addition to being classified as a female face.

An additional interesting point that can be observed for figure 3 is that the reconstruction with the eigenvectors with large eigenvalues contains mostly low-frequency information. In contrast, the reconstruction using the eigenvectors with small eigenvalues contains mostly high-frequency information. This suggests some similarities between the PCA of information contained in facial patterns and the spatial frequency analysis (SFA) of face images (Bowns and Morgan 1993; Watt

1992). These two approaches can be seen as ways of representing a set of objects (ie face images) by their projection on an orthogonal basis (ie eigenvectors for the PCA approach, and complex exponentials—or cosine and imaginery sine— for the SFA approach).

The two approaches are similar in that the variance (ie the eigenvalue) of an eigenvector can be interpreted as its energy, and since low spatial frequencies, in natural images like faces, carry more energy than high spatial frequencies, eigenvectors with large eigenvalues tend to capture more low-frequency information than eigenvectors with small eigenvalues. An important difference between these approaches, however, comes from the fact that the basis functions of the SFA are sample independent (ie SFA operates at the level of individual faces) whereas the basis functions of the PCA are sample dependent (ie PCA operates on a set of faces) since they are defined as maximizing the variance of the projections of the objects (ie face images) to be represented. Hence, these two approaches, though related, are not equivalent as is illustrated by figure 4.

Figures 4b and 4d show a Caucasian face reconstructed with, respectively, either the first 30 eigenvectors or all but the first 30 eigenvectors of an autoassociative memory trained on a sample of 160 Caucasian faces. As a comparison 4c and 4e show the same face after, respectively, low-pass filtering (ie only the low frequencies are preserved) and high-pass filtering (ie only the high frequencies are preserved). It is clear from this example that filtering a *learned* face through the autoassociative memory (ie PCA approach) is somewhat similar to filtering it through spatial frequencies filters (ie SFA approach). Figure 4f–4j shows how the two approaches differ. Figures 4g and 4i, respectively, show a Japanese face reconstructed with either the first 30 eigenvectors or all but the first 30 eigenvectors of an autoassociative memory trained on a sample of 160 Caucasian faces. Figures 4h and 4j show the same face after low-pass filtering and high-pass filtering, respectively. It is clear from this second example that in addition to acting as spatial-frequency filters, the eigenvectors distort *new* faces as an inverse proportion of their similarity with the faces corresponding to the sample from which the eigenvectors were extracted.

FIGURE 3. Illustration of the different kinds of facial information conveyed by different ranges of eigenvectors. From left to right (a) the original face; (b) its reconstruction with the first 20 eigenvectors ($r = .090$, $r^2 = 0.81$); (c) its reconstruction with eigenvectors 20 to 100 ($r = 0.42$, $r^2 = 0.17$).

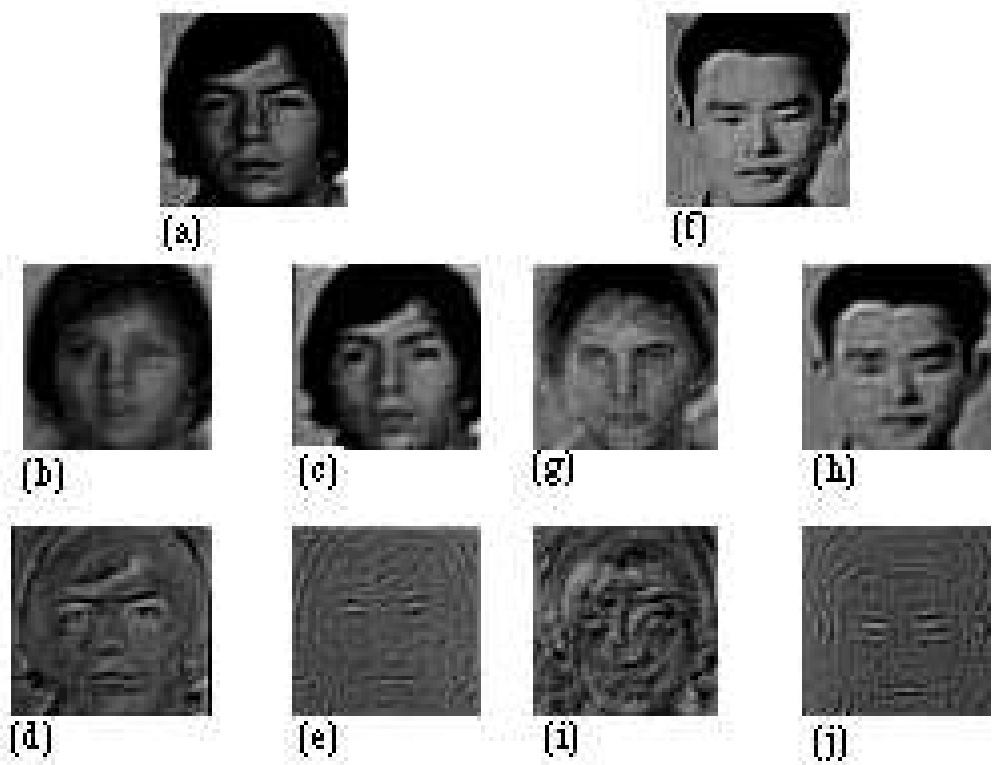

FIGURE 4. Comparison of principal-component analysis and spatial-frequency analysis of face images. (a) A learned Caucasian face; the face was reconstructed with (b) the first 30 or (d) all but the first 30 eigenvectors of an autoassociative memory (ie pixel cross-product matrix) trained with 160 Caucasian faces; (c) low-pass and (e) high-pass filtering of the same face. (f) a new Japanese face; the face was reconstructed with (g) the first 30 or (i) all but the first 30 eigenvectors of an autoassociative memory (ie pixel cross-product matrix) trained with 160 Caucasian fces; (h) low-pass and (j) high-pass filtering of the same face. Note that in addition to acting as a spatial-frequency filter, the eigenvectors distort the new face in proportion to its difference from the set of learned faces.

Another way of looking at the relationship between the PCA and the SFA approaches is to make the SFA sample dependent. This can be accomplished by considering each pixel as a receptor, and by applying a discrete Fourier transform (DFT, cf Brigham 1988) to the values of the pixels across the 160 faces of the sample, leading to 33975 DFTs (1 for each pixel position), each of them having 160 elements (1 for each face). Then it is possible to display the different components of the 33975 DFTs graphically as it is done for the eigenvectors. As an illustration, figure 5 displays the first three eigenvectors extracted from the cross-product matrix of the 160 faces of the sample (top panels) and the first three components of the 33975 DFTs. A first point to note is that the components of the DFT appear face-like and somewhat similar to the eigenvectors. Precisely, the 33975-dimensional vector collecting the magnitude of the DC components (bottom left panel) gives the average value for each pixel. Because the first eigenvector is very close to the average vector for faces ($r^2 = 0.98$, Valentin

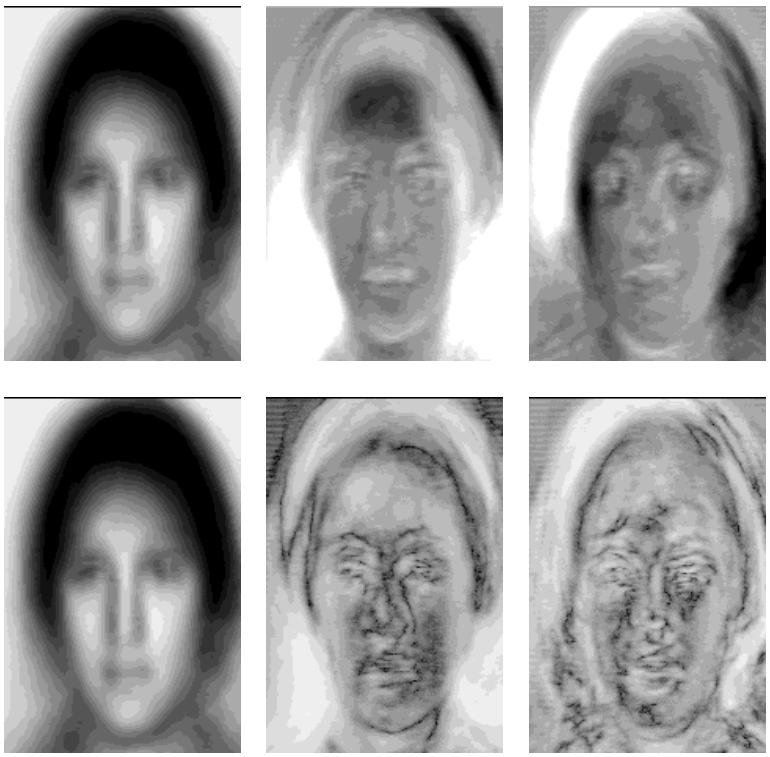

FIGURE 5. The first 3 eigenvectors of an autoassociative memory trained on a sample of 160 faces (top panels) and the magnitude of the first 3 components of a series of 33975 Discrete Fourier Transforms (1 for each pixel describing the faces) applied across the same sample of 160 faces (bottom panels). The first eigenvector is almost equivalent to the DC components of the Fourier transform ($r^2 = .98$). They both are very close to the average of the faces in the sample. The similarity between the subsequent eigenvectors and components decreases progressively ($r^2 = 0.44$ and $r^2 = 0.19$).

et al, in press), the DC component and the first eigenvector are almost identical. The second and third components of the DFTs are still substantially correlated with their equivalent eigenvectors but much less so ($r^2 = 0.44$, and $r^2 = 0.19$, respectively). This indicates again that even though PCA and SFA are related they are not equivalent.

A third approach, more usual in the domain of image processing (cf Gonzalez and Woods 1992; Jain 1989; Pratt 1991), could be to compare the SFA of an image with the PCA analysis of this image only. This is referred to as the Hotelling, Karnuhen-Loève, or singular value decomposition (SVD) approach in the literature, and the relationship of these two transforms is discussed in the sources mentioned.

To continue our analysis of the preprocessing step we examined the utility and generalizability of low-dimensional eigenvector representations of faces for supporting a gender-classification task. We wished particularly to use this technique to explore the usefulness of different eigenvector information for generalizing to novel faces. Whereas O'Toole et al (1993) looked at the utility of individual eigenvectors for predicting the sex of a face that was learned by the model, in the present study we compare the quality of *generalizable* gender information by using learned and new faces reconstructed

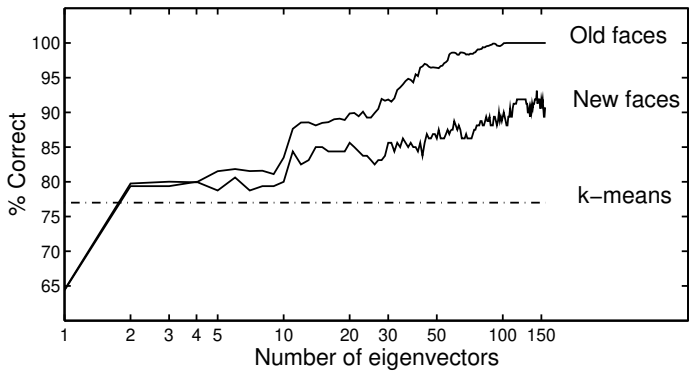

FIGURE 6. Percentage of correct gender classification as a function of the number of eigenvectors (logarithmic scale) used to represent the faces (the eigenvectors are ordered according to their eigenvalue). Performance for the old faces is averaged across trials.

with different subsets of eigenvectors. We started with the first eigenvector, which explains the largest proportion of variance, and then proceeded to supplement this representation by using additional eigenvectors successively explaining larger proportions of variance in face representations.

Since similar performance was obtained with model 1 and model 2, and since the perceptron is conceptually simpler, in what follows, we report only the results attained by model 1. The stimuli and procedure were the same as those used in the previous simulation, except that progressively increasing cumulative ranges of eigenvectors (from 1 to 159) were used to reconstruct the faces prior to categorization.

Figure 6 displays the proportion of correct gender classifications as a function of the number of eigenvectors used to represent the faces. As previously, performance for the old faces was averaged across the trials. Note that the eigenvectors are ordered according to their eigenvalues. 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. It can be seen from this figure that up to the 20th eigenvector, the accuracy of categorization increases with the number of eigenvectors both for old and for new faces. When more eigenvectors were used to categorize old faces, performance increased smoothly until a perfect categorization score was obtained with 105 eigenvectors. By contrast, increasing the number of eigenvectors beyond 20 did not significantly improve the categorization accuracy for new faces, which slowly increased from 85% to 90%.

The fact that adding eigenvectors with relatively small eigenvalues improves the performance for the faces learned by the memory, but not for new faces, suggests that there are two types of information available in the eigenvectors. The first type, contained in the first 20 eigenvectors (ie the ones with the largest eigenvalues) can be generalized to categorize new faces. The second type of information, captured by the eigenvectors with smaller eigenvalues, is specific to the faces that have been learned by the model.

Another point worth noting is that both for old and for new faces, when the second eigenvector is added to the first one, the performance of the model increases dramatically (from 64.46% to 79.74% for the old faces and from 64.37% to 78.75% for the new faces). This is consistent with previous data indicating that the second eigenvector extracted from a memory matrix, made of an equal number of male and female faces, plays an important role in the determination of the sexual appearance of the faces in the set (O'Toole et al 1993; Valentin et al, in press). This is particularly intriguing since, when graphically displayed, the second eigenvector does not seem to capture any perceptual

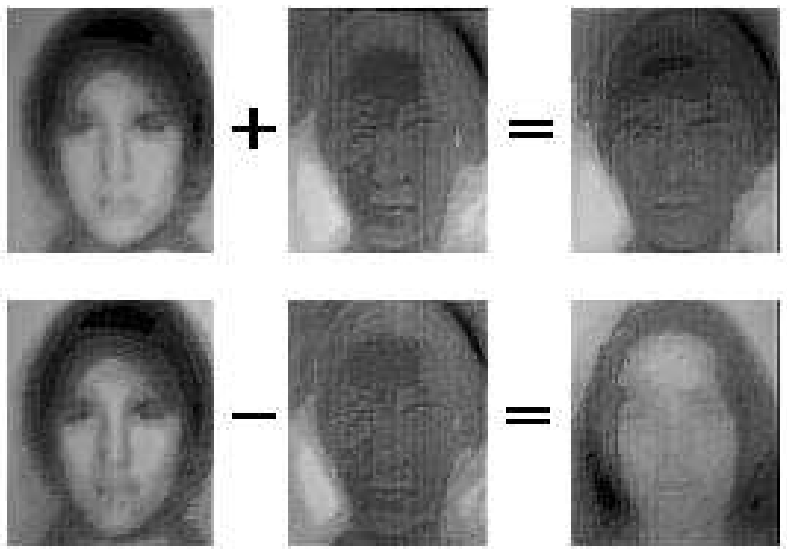

FIGURE 7. Adding the first eigenvector to the second eigenvector creates a masculine-looking face. Subtracting the second eigenvector from the first eigenvector creates a feminine-looking face [from Valentin et al (1994b), with authorization of the authors].

information relative to the sex of the faces. However, O'Toole et al demonstrated that when the second eigenvector is added to the first eigenvector it produces a masculine-looking face, but when it is subtracted from the first one, it produces a feminine-looking face (cf figure 7). Note that the sign of the eigenvectors is arbitrary.

Another way of exploring the usefulness of the information captured by the eigenvectors with large eigenvalues in categorizing faces is to project the faces onto the space defined by those eigenvectors. Since the first two eigenvectors seem to capture most of the information relative to the sexual appearance of faces, we looked at the projections of the faces onto those two eigenvectors. Figure 8 is a graphical representation of the projection of male ($\circ$) and female ($*$) faces onto the first two eigenvectors extracted from the complete set of faces. It can be seen from this figure that all the faces are positively correlated with the first eigenvector (ie they all project on the positive part of the horizontal axis), and the projections of female faces tend to cluster onto the negative part of the second eigenvector, whereas the projections of male faces tend to cluster onto the positive part of this eigenvector (as previously noted, the signs of the projections are completely arbitrary). In other words, female and male faces tend to have opposite weights on the second eigenvector. These results indicate that the second eigenvector tends to separate male faces from female faces. The first eigenvector represents the face category in general and could be used to categorize faces as opposed to other object categories or to detect a face in an image.

To examine the psychological relevance of the information contained in the second eigenvector, we attempted to manipulate the sexual appearance of faces by changing their weights on this eigenvector. Figure 9 represents a male and a female face after modification of the weight on the second eigenvector. The original faces are presented in the middle panels, the "feminized" faces (ie with a negative weight twice the absolute value of the original weight) on the left panels and the "masculinized" faces (ie with a positive weight twice the absolute value of the original weight) on the right

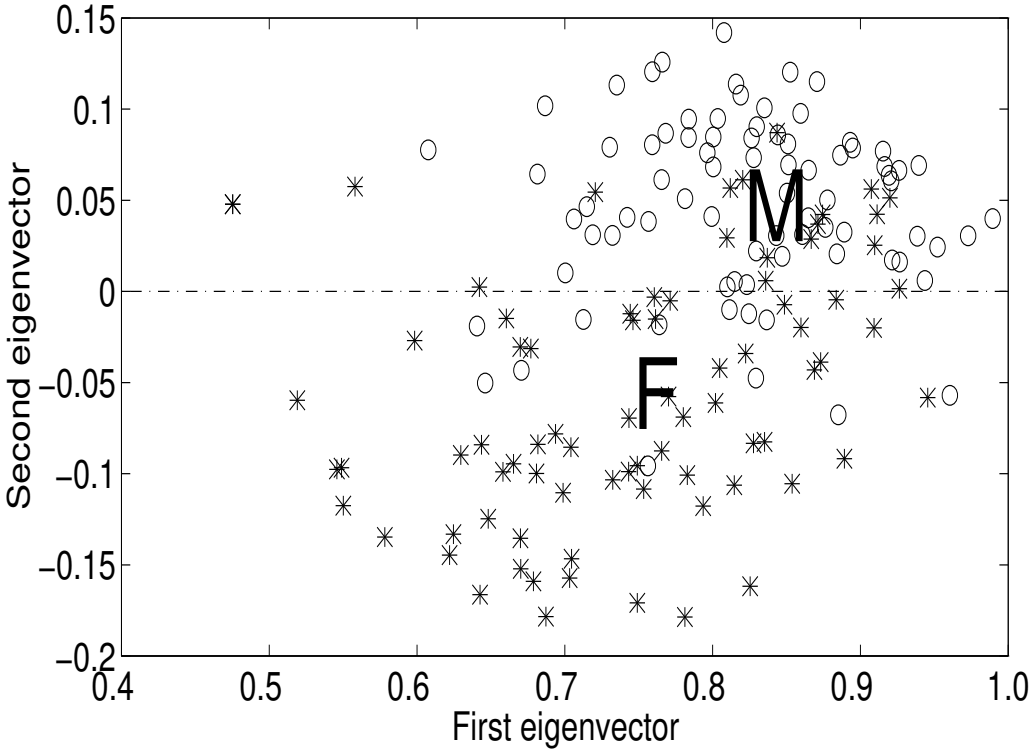

FIGURE 8. Projection of female (*) and male (o) faces onto the space determined by the first two eigenvectors of a face-autoassociative memory (the eigenvectors are normalized to unity). It can be seen that the projections of the male faces tend to cluster (M) on the positive part of the second eigenvector and the female faces to cluster (F) on the negative part.

panels. Clearly, the manipulation of the weights on the second eigenvector modifies the apparent femininity/masculinity of the faces: the female in the upper right panel looks more "boyish" and the male in the lower left more "girlish." However, further work is necessary to test if such manipulations would affect subject performance on a categorization-rating task.

Last, in order to analyze more precisely the perceptual information captured by the second eigenvector, we examined the pixels that contribute strongly to this eigenvector (ie the ones with the larger absolute values). An easy way of doing that is to display the pixels with values greater than some arbitrary threshold (for a more detailed explanation, see O'Toole and Abdi 1989). Here we used the mean of the absolute pixel values for this particular eigenvector as a threshold. To represent the pixels that contribute strongly to the second eigenvector, we used two different display schemes corresponding respectively to the left and right panels of figure 10. In the left panel, the pixels contributing positively are displayed in black and the pixels contributing negatively in white. In the right panel, the pixels with negative contributions are displayed in black, and the pixels with positive contributions in white (note that the left panel is the inverse image of the right one). In both panels, the gray areas represent pixels that do not contribute, above the threshold, to the second

FIGURE 9. A female and a male face after modification of the weight on the second eigenvector. The middle panels represent the original faces. The left panels represent the "feminized" faces (ie the weight is set to a negative value twice the absolute value of the original weight). The right panels represent the "masculinized" faces (ie the weight is set to a positive value twice the absolute value of the original weight).

eigenvector. These pixels, being less important for the computation of the second eigenvector, are hence less important for discriminating between male and female faces using this eigenvector. The coding scheme used in the right panel reflects the actual value of the pixels for most of the male faces and the coding scheme used in the left panel reflects the actual value of the pixels for most of the female faces (ie white indicates a low pixel intensity, black a high pixel intensity).

The major observation that can be derived from figure 10 is that the most important areas, when only the second eigenvector is used for discriminating male and female faces, are the hair, forehead, eyebrows, nose and chin area. Further examination of the intensity of the pixels composing these areas for male and female faces suggests, in agreement with Brunelli and Poggio (1992) and Burton et al (1993), that male faces tend to have a longer chin, a bigger nose, thicker eyebrows, and shorter hair than female faces. Of course, additional information useful for discriminating female and male faces is available in other areas or aspects of the faces not captured by the second eigenvector. Further study is necessary to determine the characteristics of this information.

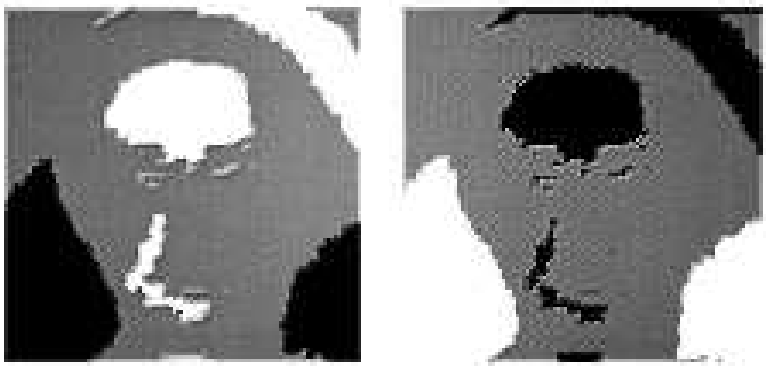

FIGURE 10. Pixels contributing strongly to the second eigenvector. In the left panel, the pixels contributing positively are displayed in black and the pixels contributing negatively are displayed in white. In the right panel, the pixels contributing negatively are displayed in black and the pixels contributing positively are displayed in white (the image in the right panel is the negative of the image presented in the left panel).

6. Role of the hair in model performance

The fact that pixels corresponding to the hair area contribute strongly to the second eigenvector suggests that the shape and length of the hair constitute an important cue for the model to separate male and female faces. An examination of the faces that were misclassified by model 2 is one way to examine further this issue. The hairstyles of the 80 male stimuli were all quite similar [5], while the 80 female stimuli showed more variance, especially in hair length. Therefore, we focus on female faces, of which 28 out of 80 had hair similar in length to the male stimuli and 52 had longer hair. Model 2 misclassified 10 female images as male; 6 of these misclassified females had hairstyles similar in length to the average male stimulus and four had hairstyles that were considerably longer than that of the average male stimulus. In other words, over 20% of the females with atypical or "short" hair were misclassified, in contrast to about 8% of the females with long hair. A Chi-square test reveals a marginally significant trend for misclassification of females with atypically short hair $\chi^2(1) = 3.14$, $p = .073$. This tendency again raises the question of the importance of the hairstyle to the performance of the model. Although retaining hair information in the stimuli provides some ecological validity, it seems important to partial out the contribution of this factor in gender prediction. For an estimation of the performance of the model without this information, we carried out a last series of simulations with information on hair shape and length omitted.

6.1. *Stimuli.* The stimuli were the same 160 faces used for the previous simulations. To discard hair-shape information the faces were first cropped as illustrated by figure 11. Then to eliminate any information relative to the length of the hair in the cropped images, the pixels contributing the most to the second eigenvector (ie with a value greater than the mean plus 1 standard deviation) were set to 0. These pixels appear in black in figure 11. As can be seen in this figure, very little information relative to hairdo remained on the face images.

6.2. *Procedure.* The procedure was the same as the one used in the first simulation for model 2 (an autoassociator followed by a perceptron). A jackknife technique was applied to our sample

[5] The stimuli were pictures taken in the 1970's and hence male hair typically ranged in length from the bottom of the ear to the top of the collar.

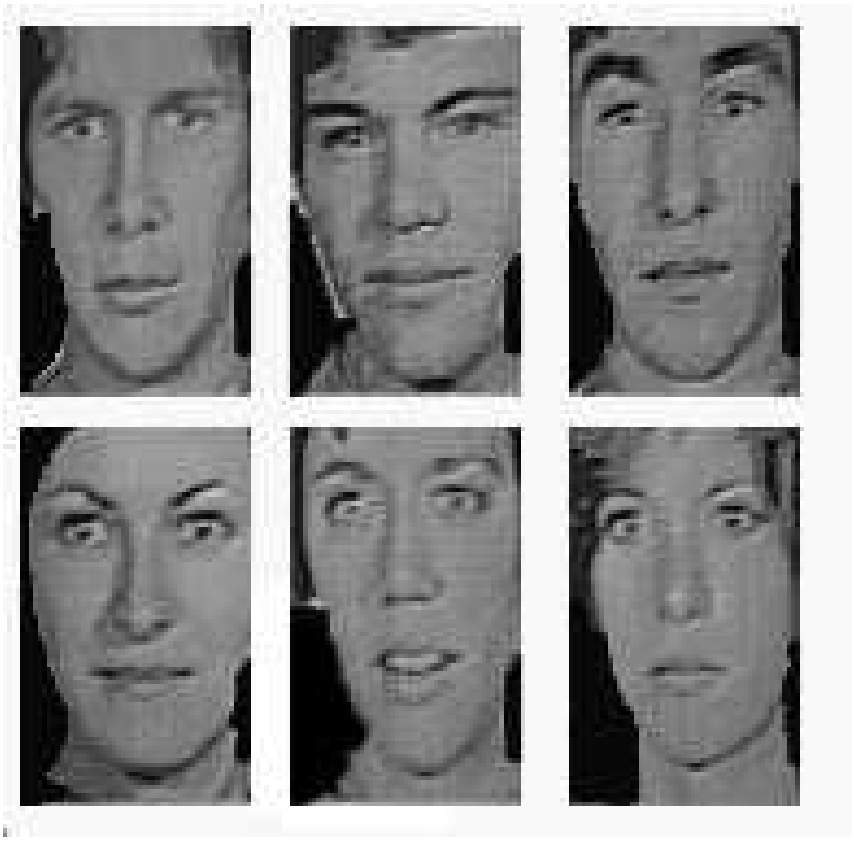

FIGURE 11. Example of stimuli after elimination of information on hair shape and length. The top panels represent 3 male faces, the bottom panels 3 female faces. Note that little information relative to the hair shape and length is preserved in the images.

of 160 cropped face images as previously explained. The faces were first pre-processed via an autoassociative memory (ie eigendecomposition) and each face was coded by a 159-dimensional vector corresponding to the values of its projections on the 159 eigenvectors (or principal components) of the autoassociative memory. The prediction of the gender of the faces (new and old) was achieved by using a perceptron. The perceptron was trained to associate 1 with female faces and 0 with male faces. In addition, to provide a baseline for classification performance with these stimuli a series of 1000 k-means classifications was performed.

6.3. *Results.* The results of this simulation showed 80% correct classification as compared with the 90% attained with the "hairy" stimuli, thus indicating some loss of gender-prediction power. However, the mean correct gender classification obtained with the k-means algorithm with the same stimuli dropped to 56%. Even though the performance of the perceptron decreased by 10% when the hair was eliminated, it still substantially outperformed the unsupervised k-means. This indicates that while the k-means classification relies quite heavily on the hair (performance barely above chance), the supervised model (autoassociator followed by a perceptron) is able to employ other types of information.

A last point worth noting is that when the hair information was omitted, no difference in k-mean classification performance was observed between male and female faces (55% for the males vs

57% for the females, with a median .53 and .55 respectively). This last point demonstrates again the importance of the hair in the unsupervised k-means algorithm. Recall that when k-means was applied to stimuli that included the hair, prediction was better for male than for female faces (87% *vs.* 66%, respectively)

Thus, in conclusion, it appears that although the hair makes some contribution to the classification process, it is not the sole contributor; there are other factors influencing the classification, such as global shape.

7. Conclusion

Our major purpose was to investigate the ability of a low-level representation to produce sex-classification performance comparable to that achieved by measurement-based models. To do this we tested several supervised models, concentrating on their ability to generalize to new faces, against a baseline established by an unsupervised algorithm. A simple perceptron and an RBF network yielded equivalent best performances (about 90%) when preceded by an eigendecomposition preprocessor. These results are similar to those of the measurement-based method, require no a priori selection of features, and also provide a byproduct (ie eigenvectors) for further analysis. To assess the importance of hair length and shape to the classification performance of the models, in a final simulation this information was omitted from the stimuli: the results showed that hair style, although contributory, is not of sole importance.

In addition to enabling simple neural networks to categorize faces according to their gender, the PCA/autoassociator preprocessing of the face images constitutes an interesting tool for quantifying or analyzing the information determining the sexual appearance of faces. Specifically, the preceding analysis indicates that the model exploited two kinds of information in learning to categorize the faces by gender. The first one, conveyed by the eigenvectors with larger eigenvalues, can be generalized to new faces. It encompasses general information relative to the global shape of the face, hairstyle, as well as broad areas such as the nose, eyebrows and chin regions. The second type of information, conveyed by the eigenvectors with smaller eigenvalues, did not generalize. From the data we have presented, it is not possible to determine, precisely, the nature of this latter kind of information. At least two interpretations are possible. First, it may be that this latter information is idiosyncratic gender information. In other words, it is possible that this information is useful for determining the masculine/feminine appearance of a single or small number of faces in the training set, but is not generally useful for determining the sexual appearance of most or all faces. The second possibility is that the information is not really useful for determining the sexual appearance of the faces (for human perceivers), but that the model was nonetheless able to use the information to assign the face to the "correct" category. This latter possibility occurs owing to the "overfitting" problem, which gives the model a very large number of predictors with which to work. A third possibility, of course, is that both of these types of information contribute to the classification ability of the model for learned faces and neither generalizes to the new faces. Deciding between these hypotheses, however, is a matter for further study.

Acknowledgements. Thanks are due to June Chance and Al Goldstein for providing the faces used in the simulations and to M Morgan and an anonymous reviewer for helpful comments on a previous draft of this paper.

References

[1] Abdi H, 1988 "A generalized approach for connectionist auto-associative memories: interpretation, implications and illustration for face processing", In *Artificial intelligence and cognitive sciences* Ed J Demongeot (Manchester: Manchester University Press) pp 149–164

[2] Abdi H, 1994a *Les réseaux de neurones* (Grenoble: Presses Universitaires de Grenoble)

[3] Abdi H, 1994b "A neural network primer" *Journal of Biological Systems* **2** 247–281

[4] Abdi H, Valentin, D 1994 "Modèles neuronaux, connexionistes et numériques pour la mémoire de visages" *Psychologie Française* **39** 375–391

[5] Anderson J A, Mozer M C, 1981 "Categorization and selective neurons" In *Parallel models of associative memory* Eds G E Hinton, J A Anderson (Hillsdale: Erlbaum) pp 213–236

[6] Anderson JA, Silverstein JW, Ritz SA, Jones, RS, 1977 "Distinctive features, categorical perception, and probability learning: Some applications of a neural model" *Psychological Review* **84** 413–451

[7] Bowns L, Morgan MJ, 1993 "Facial features and axis of symmetry extracted using natural orientation information" *Biological Cybernetics* **70** 137–144

[8] Brigham EO, 1988 *The fast Fourier transform and its applications* (Englewood Cliffs NJ: Prentice-Hall)

[9] 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

[10] Bruce V, Burton A M, Dench N, Hanna E, Healey P, Mason O, Coombes A, Fright R, Linney A, 1993 "Sex discrimination: How do we tell the difference between male and female faces?" *Perception* **22** 131–152

[11] Brunelli R, Poggio T, 1992 "HyperBF Networks for sex classification" *Proceedings of the Image Understanding Workshop* DARPA San Diego January 1992

[12] Brunelli R, Poggio T, 1993 "Face recognition: Features versus templates" *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15** 1041–1052

[13] Burton A M, Bruce V, Dench N, 1993 "What's the difference between men and women? Evidence from facial measurement" *Perception* **22** 153–176

[14] Cottrell G W, Fleming M K, 1990 "Face recognition using unsupervised feature extraction" In *Proceedings of the International Neural Network Conference* (Paris, France. Dordrecht: Kluwer) pp 322–325

[15] Cottrell G W, Metcalfe J, 1991 "EMPATH: Face, gender and emotion recognition using holons" In *Advances in neural information processing systems 3* Eds R P Lippman, J Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 564–571

[16] Efron B, 1979. "Bootstrap methods: Another look at the jackknife" *Annals of Statistics* **7** 1–26

[17] Golomb B A, Lawrence D T, Sejnowski T J, 1991 "Sexnet: a neural network identifies sex from human face". In *Advances in neural information processing systems 3* Eds R P Lippman, J Moody, D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 572–577

[18] Gonzalez R C, Woods RE, 1992 *Digital Image Processing* (Reading MA: Addison Wesley)

[19] Gray M S, Lawrence D T, Golomb B A, Sejnowski T J (submitted) "A perceptron reveals the face of sex"

[20] Jain A K, 1989 *Fundamentals of Digital Image Processing* (Englewood Cliffs NJ: Prentice-Hall)

[21] Nadler M, Smith E P, 1993 *Pattern recognition engineering* (New-York: Wiley) pp 298–300

[22] Michelli C A, 1986 "Interpolation of scattered data: Distance matrices and conditionally positive definite functions" *Constructive Approximation* **2** 11–22

[23] O'Toole A J, Abdi H, 1989 "Connectionist approaches to visually based feature extraction" In *Advances in cognitive psychology (Vol 2)* Ed G Tiberghien (London: John Wiley)

[24] 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 *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* Eds K J Hammond D Gentner (Hillsdale N J: Erlbaum) pp 847–851

[25] 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

[26] O'Toole A J, Deffenbacher K A, Valentin D, Abdi H, 1994 "Structural aspects of face recognition and the other race effect" *Memory and Cognition* **22** 208–224

[27] Pratt W K, 1991 *Digital Image Processing* (New York: Wiley)

[28] Rhodes G, 1988 "Looking at faces: First-order and Second order features as determinant of facial appearances" *Perception* **17** 43–63

[29] Samal A, Iyengar P A, 1992 "Automatic recognition and analysis of human faces and facial expressions: A survey" *Pattern Recognition* **25** 65–77

[30] 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

[31] Valentin D, Abdi H, O'Toole AJ, (in press) "Principal component and neural network analyses of face images: Exploration into the nature of the information available for classifying faces by sex". In *Progress in Mathematical Psychology* Eds C Dowling, F S Roberts, P Theuns (Hillsdale NJ: Erlbaum)

[32] Valentin D, Abdi H, O'Toole AJ, 1994 "Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches" *Journal of Biological System* **2** 413–429

[33] Valentin D, Abdi H, O'Toole AJ, Cottrell GW, 1994 "Connectionist models of face processing: A survey" *Pattern Recognition* **27** 1209–1230

[34] Van de Geer, JP, 1971 *Introduction to Multivariate Analysis for the Social Sciences* (San Francisco: Freeman)
[35] Watt RJ, 1992 "Faces and vision". In *Processing Images of Faces* Eds V Bruce, M Burton (Norwood NJ: Ablex Publishing Corporation)

APPENDIX 1

Autoassociative memory, eigendecomposition, and singular-value decomposition

To store a set of faces in an autoassociative memory, each face is first digitized and the rows of the pixel image concatenated to form an $I \times 1$ vector $\mathbf{x}_k$ whose I components represent the gray level of the I pixels. In a neural network implementation, these components represent the activation of the input units (ie cells). For convenience, the vectors $\mathbf{x}_k$ are assumed to be normalized so that $\mathbf{x}_k^T \mathbf{x}_k = 1$ (with $\mathbf{x}_k^T$ denoting the transpose of $\mathbf{x}_k$). The set of K faces to be stored in the memory is represented by an $I \times K$ matrix $\mathbf{X}$ in which the kth column is equal to $\mathbf{x}_k$. The autoassociative memory (or weight matrix) is represented by an $I \times I$ matrix $\mathbf{W}$. The values in the weight matrix correspond to the connection strengths between the units of the memory.
The faces are stored in the memory by changing the strength of the connections between units. This can be done using a simple Hebbian learning rule:

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

Recall of a given stimulus $\mathbf{x}_k$ is given by:

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

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

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

where $\|\mathbf{x}_k\|$ is the euclidean norm of the vector $\mathbf{x}_k$ [ie $\|\mathbf{x}_k\| = \sqrt{\mathbf{x}_k^T \mathbf{x}_k}$]. A cosine of 1 indicates a perfect reconstruction of the stimulus.
When the stimulus set is composed of non-orthogonal stimuli, the associator does not perfectly reconstruct the stimuli that are stored. On the other hand, some new patterns are perfectly reconstructed, creating, in a way, the equivalent of a "false alarm" or "false recognition." These patterns are defined by the equation:

$$\mathbf{W}\mathbf{u}_r = \lambda_r \mathbf{u}_r \qquad \text{with} \quad \mathbf{u}_r^T \mathbf{u}_r = 1 \tag{4}$$

where $\mathbf{u}_r$ denotes the rth eigenvector of $\mathbf{W}$, and λ_r the eigenvalue associated with the rth eigenvector.
From equation(1) it can be seen that the matrix $\mathbf{W}$ is equivalent to a cross-product matrix, and hence is positive semidefinite (ie all its eigenvalues are positive or zero, and its eigenvectors are real valued and pairwise orthogonal). Consequently, $\mathbf{W}$ can be reconstructed as a weighted sum of its eigenvectors:

$$\mathbf{W} = \sum_{r=1}^{R} \lambda_r \mathbf{u}_r \mathbf{u}_r^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \qquad \text{with} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \tag{5}$$

where $\mathbf{I}$ stands for the identity matrix, $\mathbf{\Lambda}$ represents the diagonal matrix of eigenvalues, and R is the rank of the matrix $\mathbf{W}$ (ie the number of its non-zero eigenvalues). The eigenvectors in $\mathbf{U}$ are usually ordered according to their eigenvalues from large to small. This formulation makes clear the close relationship between the classical linear autoassociator and some techniques used in multivariate statistical analysis. Specifically, using an autoassociative memory to store and recall a

set of objects is equivalent to performing a PCA on the cross-product matrix of the feature set describing these objects (cf Anderson et al 1977).
A final point worth noting is that the eigenvectors and eigenvalues of the weight matrix $\mathbf{W}$ can be obtained directly using the singular-value decomposition (SVD) of the original matrix of stimuli $\mathbf{X}$. Formally:

(6) $$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{V}^T \qquad \text{with } \mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{U} = \mathbf{I}$$

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 root of the eigenvalues of $\mathbf{X}\mathbf{X}^T$ or $\mathbf{X}^T\mathbf{X}$ (they are the same). The projections, $\mathbf{G}$, of the K stimuli of the training set on the R eigenvectors of the weight matrix can be found as:

(7) $$\mathbf{G} = \mathbf{X}^T\mathbf{U} = \mathbf{V}\mathbf{\Delta} \ .$$

Often, for convenience the projections are normalized to 1, and hence

(8) $$\mathbf{G}_{\text{norm}} = \mathbf{G}\mathbf{\Delta}^{-1} = \mathbf{V} \ .$$

Likewise, the projections of a set of K' new stimuli (ie not learned by the memory), $\mathbf{X}_{\text{new}}$, on the eigenvectors of $\mathbf{W}$ can be computed as:

(9) $$\mathbf{G}_{\text{new}} = \mathbf{X}_{\text{new}}^T\mathbf{U} \ .$$

In order to improve the storage capacity of an autoassociative memory, most applications involve the Widrow-Hoff learning rule. The Widrow-Hoff learning rule corrects the difference between the response of the system and the expected response by changing iteratively the weights in matrix $\mathbf{W}$ as follows:

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

where η is a small positive constant and t is the iteration step. The Widrow-Hoff learning rule can also be analyzed in terms of the eigenvectors and eigenvalues of $\mathbf{W}$ (Abdi, 1994a). Specifically, $\mathbf{W}$ at time t can be expressed as:

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

With a positive learning constant η smaller than $2\lambda_{\max}^{-1}$ ($\lambda_{\max}$ being the largest eigenvalue) this procedure converges towards

(12) $$\mathbf{W}_{[\infty]} = \mathbf{U}\mathbf{U}^T$$

which indicates that using the Widrow-Hoff error correction learning rule amounts to equalizing all the eigenvalues of $\mathbf{W}$ (ie to sphericizing the weight matrix).

APPENDIX 2

The perceptron

The perceptron is a simple classifier. In its basic version, it is composed of one input layer (sometimes referred to as the retinae of the perceptron), and one binary output cell. Its purpose is to classify a set of objects into two classes.
If the objects to be classified are represented by $I \times 1$ vectors denoted $\mathbf{x}_k$, the goal of the perceptron is to find a set of I weights w_i stored in vector $\mathbf{w}$ such that

(13) $$\mathbf{w}^T\mathbf{x}_k > 0 \quad \text{for } k \text{ belonging to the first category}$$

and

(14) $$\mathbf{w}^T\mathbf{x}_k \leq 0 \quad \text{for } k \text{ belonging to the second category.}$$

Suppose that the K objects to be classified are stored in an $I \times K$ matrix denoted $\mathbf{X}$. In the case of two groups, the optimum solution equivalent to discriminant analysis is found (cf Van de Geer 1971) as

(15) $\mathbf{w}^T = \mathbf{y}^T\mathbf{X}^+$

where $\mathbf{y}$ is the $K \times 1$ target vector (with $y_k = 1$ for k belonging to the first group and with $y_k = 0$ for k belonging to the second group), and $\mathbf{X}^+$ is the Moore-Penrose pseudo-inverse of $\mathbf{X}$. Using the notation of appendix 1,

(16) $\mathbf{X}^+ = \mathbf{V}\mathbf{\Delta}^{-1}\mathbf{U}^T$.

APPENDIX 3

Radial basis function network

RBF networks are two-layer feed-forward networks in which the input vectors $\mathbf{x}_k$ are propagated to a layer of hidden units. Each hidden unit computes the radial basis function ϕ of the distance from the input vectors $\mathbf{x}_k$ to its center $\mathbf{c}$:

(17) $\mathbf{h}_i = \phi(\|\mathbf{x}_k - \mathbf{c}_i\|)$

where $\mathbf{h}_i$ is the output of the ith hidden unit and $\|.\|$ denotes the euclidean norm. A variety of RBFs can be chosen, the most popular one is probably the Gaussian function:

(18) $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-x^2/2\sigma^2\}$.

Note that the constant term ($\sqrt{2\pi\sigma^2}$) is used so that the integral of $\phi(x)$ is normalized to 1 and can be dropped from equation (18) without any problem.
The outputs of the hidden units are then propagated to the output units of the network which integrates them in the following manner:

(19) $\mathbf{o}_j = \sum_i w_{i,j}\mathbf{h}_i = \sum_i w_{i,j}\phi\{\|\mathbf{x}_k - \mathbf{c}_i\|\}$

where $w_{i,j}$ represents the strength of the connection between hidden unit i and output unit j. Learning is achieved by iteratively changing the weights $w_{i,j}$ so as to minimize the error between the actual and the desired output. Rewriting equation (19) with a matricial notation makes clear that the optimal weights can be found using a least-squares approximation (cf Abdi 1994b). In brief, if $\mathbf{D}$ denotes the matrix of euclidean distance between each stimulus and each center and $\mathbf{O}$ the output matrix, equation (19) can be expressed as:

(20) $\mathbf{O} \approx [\phi(\mathbf{D})]\mathbf{W}$

with the function ϕ being applied elementwise to the elements of $\mathbf{D}$. If the matrix $\phi(\mathbf{D})$ is square, and non-singular, the solution for the matrix $\mathbf{W}$ is obviously

(21) $\mathbf{W} = [\phi(\mathbf{D})]^{-1}\mathbf{O}$

where $[\phi(\mathbf{D})]^{-1}$ is the inverse of matrix $[\phi(\mathbf{D})]$. If $\phi(\mathbf{D})$ is singular or rectangular, a least-squares approximation is given by

(22) $\mathbf{W} \approx [\phi(\mathbf{D})]^{+}\mathbf{O}$

where $[\phi(\mathbf{D})]^{+}$ represents the Moore-Penrose (or pseudo) inverse of $\phi(\mathbf{D})$ (cf appendix 2).