Cognitive Science 25 (2001) 819–838

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Classifying adults' and children's faces by sex: computational investigations of subcategorical feature encoding

Yi D. Cheng, Alice J. O'Toole*, Hervé Abdi

The University of Texas at Dallas, School of Human Development, GR 4.1, Cognition & Neuroscience Program, Richardson, TX 75083-0688, USA

Abstract

The faces of both adults and children can be classified accurately by sex, even in the absence of sex-stereotyped social cues such as hair and clothing (Wild et al., 2000). Although much is known from psychological and computational studies about the information that supports sex classification for adults' faces, children's faces have been much less studied. The purpose of the present study was to quantify and compare the information available in adults' versus children's faces for sex classification and to test alternative theories of how human observers distinguish male and female faces for these different age groups. We implemented four computational/neural network models of this task that differed in terms of the age categories from which the sex classification features were derived. Two of the four strategies replicated the advantage for classifying adults' faces found in previous work. To determine which of these strategies was a better model of human performance, we compared the performance of the two models with that of human subjects at the level of individual faces. The results suggest that humans judge the sex of adults' and children's faces using feature sets derived from the appropriate face age category, rather than applying features derived from another age category or from a combination of age categories. © 2001 Cognitive Science Society, Inc. All rights reserved.

1. Introduction

Adults, children, and infants discriminate between male and female faces quickly and accurately, (e.g., Burton, Bruce & Dench, 1993; Fagot & Leinbach, 1993; Intons-Peterson,

* Corresponding author. Tel.: +1-972-883-2486; fax: +1-972-883-2491.
E-mail address: otoole@utdallas.edu (A. O'Toole).

PII: S0364-0213(01)00054-4

1988). Our ability to categorize faces by sex is due presumably to the fact that the sex of a face can be determined from reliable cues or features which we learn to apply to the task. Within the general category of faces, however, there are a number of subcategories, including race and age. Despite the relatively profound face structure differences characterizing different facial subcategories, the sex of a face is a categorical distinction that remains relevant across the subcategories.

An example of the perceptual implications of multiple coexisting subcategories of faces in the context of perceiver experience can be found in O'Toole, Peterson and Deffenbacher (1996). They demonstrated an "other-race" effect in a sex-classification task. Specifically, O'Toole et al. found that both Caucasian and Asian observers showed an own-race accuracy advantage when classifying faces by sex. This finding suggests that there may be differences in the features that specify sex for faces of different races and/or in our ability to extract these features efficiently from different categories of faces. Thus, sex categorization may interact with other subcategorical distinctions in faces, such as race, and possibly also with the experience of the observer.

Considering the subcategory of face age, Wild, Barrett, Spence, O'Toole, Cheng and Brooke (2000) investigated first-graders', third-graders',[1] and adults' ability to classify adults' and children's faces by sex in the absence of sex-stereotyped cues, such as those found in hairstyle and clothing. Previous studies have shown that when these cues are present, infants as young as 9 months of age are able to categorize pictures of males and females (Fagot & Leinbach, 1993; Cornell, 1974). Without sex-stereotyped cues, however, the task becomes more difficult. Wild et al. trimmed the faces of 7 to 10 year old children and young adults to exclude these cues. They found that all observers categorized the adults' faces accurately, but that only the third graders and adults categorized the children's faces at above-chance levels of accuracy. In addition, both children and adults classified adults' faces more accurately than children's faces. This result suggests that the sex information in children's faces may be less reliable or different than the sex information in adults' faces. In addition, even though adults' faces were categorized by sex more accurately than children's faces, adults' and children's faces were *recognized* equally accurately (Wild et al., 2000). This finding indicates that children's faces are not generally less informative than adults' faces, but rather, are less informative on the dimension of sex.

The psychological finding that children's faces are categorized by sex less accurately than adults' faces is consistent with anthropological measures of male and female children and adult faces. These data indicate that there are a number of skeletal structure differences between the faces of male and female adults (Enlow, 1982). For example, the craniofacial shape of adult males tends to be longer and less round than the shape of adult females. Additionally, on average, adult males have a larger nose and a broader forehead (Enlow, 1982). In contrast, Enlow claims that the faces of boys and girls before puberty are essentially comparable. So according to the anthropological literature, skeletal face structure cannot be used to distinguish between the faces of boys and girls.

From a perceptual standpoint, nonetheless, it seems clear that there is sufficient information in the faces of children to enable accurate sex classification. There is evidence for this claim both from the behavioral results of Wild et al. (2000) and from a morphing technique

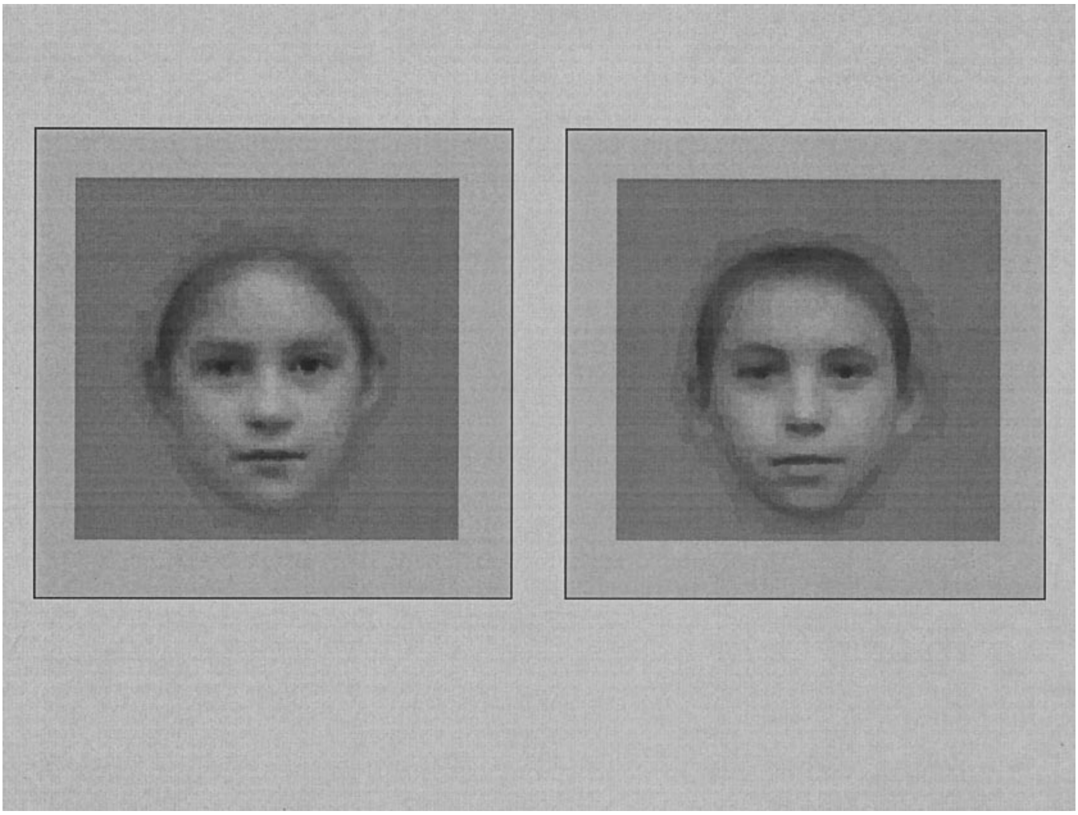


Fig. 1. Girl (left) and boy (right) prototypes created by morphed averaging.

developed by Yamaguchi, Hirukawa and Kanazawa (1995) to show the global structure differences between Japanese male and female faces. They created prototypical male and female faces by morphing together pairs of same-sex faces. They continued by morphing pairs of morphs together until a male and a female prototype were created. Wild et al. (2000) applied this technique to children's faces to obtain a prototypical boy face and a prototypical girl face (see Fig. 1). The prototypes clearly show a male-female difference that seems readily detectable.

Given the lack of skeletal differences between boys' and girls' faces, and the finding of accurate classification and distinguishable prototypes of human observers, how do we distinguish boys' and girls' faces? How does this relate to the way we distinguish men's and women's faces? A prerequisite to addressing these questions is to determine how the sex-linked information in children's faces relates to the sex-linked information in adults' faces. The information in children's faces may be a subset of the information in adults' faces, with perhaps only hormonally-based differences in tissue and fat rather than both hormonal and skeletal structure differences. It is possible also that the information in adults' faces is an exaggerated or caricatured version of the information in children's faces. At the extreme, it is possible that the information for determining the sex of children's faces is completely different than the information in adults' faces.

Computational modeling is an effective tool for analyzing the quality and nature of the sex information in faces. Although there have been numerous computational and neural network studies of sex classification with adults' faces (for reviews, see Valentin, Abdi, Edelman & O'Toole, 1997; O'Toole, Vetter, Volz & Salter, 1997), there are no analogous studies of children's faces. In this paper, we used computational learning models to compare the quality and nature of the sex information in adults' versus children's faces. We did this by manipulating the training and testing sets of the computational models in ways we will describe shortly. We were then able to compare the performance of models implemented in different ways with the performance of human subjects on the same task and with the same faces.

The primary challenge in designing computational models that are useful for comparing the sex information in adults' and children's faces is to formulate training and test sets that can be used to compare the alternative hypotheses. To the best of our knowledge, this study represents the first attempt to apply computational models to the analysis of children's faces. As such, for children's faces there is no literature to guide the design of the models. We have, therefore, formulated our models along the most basic logical dimensions. If the sex information in adults' and children's faces is similar, one would expect that the information learned from either set of faces would generalize well to the other set of faces. If, on the other hand, the sex information for adults' and children's faces is very different, we would expect little generalization between face sets. Finally, if the information in one set of faces is a subset or superset of the information in the other set of faces, we might expect a partial transfer of learning between the face sets.

From this logic, we implemented the following four simulations in which we varied the derived feature sets used for sex classification. In the *adult* feature strategy, a feature set for sex classification is derived from adults' faces and applied to classify both adults' and children's faces. This strategy has intuitive appeal based on the anthropological literature which suggests that the sex information in adults' face structure may be a superset of the sex information in children's faces. Thus, skeletal structure differences may emerge after puberty, complementing and enhancing the tissue and fat differences already there. The adult feature model would predict better performance for adults' faces because the derived feature set will be tailored to adults' faces. The model may also perform adequately on children's faces if at least a subset of the adult information is informative for sex classification of the children's faces. However, this strategy might rely on some features that would be irrelevant and potentially misleading when applied to children's faces.

In the *child* feature strategy, the sex classification features are derived from children's faces. Because the features are optimal for children's faces, this model is unlikely to duplicate the pattern of results from Wild et al. (2000). Notwithstanding, there is a particular situation that could lead to better performance for adults' faces, even though the sex classification features are derived from children's faces. Specifically, if the sex information in adult's faces is a caricatured or exaggerated version of the same sex information in children's faces, the model may perform more accurately for adults' faces. In other words, if the model learns the more difficult task of classifying children's faces, adults' faces might prove easier than the children's faces.

In the *combined* feature strategy, the sex classification feature set is derived from a combination of adults' and children's faces. This strategy seems reasonable because people see faces of all ages in their everyday experience. Thus, we may learn to classify faces by sex using the information encoded from a combination of faces of all ages, making a compromise between the information useful for adults' versus children's faces. In addition, this is a parsimonious strategy for making use of both the adults' and children's faces in deriving the feature set, because only one set of features is needed to encode both the adults' and children's faces. It is difficult to predict performance for this model. On the one hand, we might expect better performance for adults' faces during test because the behavioral data suggests that adults' faces are more informative than children's faces. On the other hand, although features derived from this strategy may conveniently exploit the common elements of the sex information in adults' and children's faces, some of these features may be inappropriately applied to faces of different ages.

In the *separate* feature strategy, two separate feature sets are derived from adults' and children's faces, respectively. In this case, a feature set is applied only to the appropriate age group of faces. This model instantiates the hypothesis that humans maintain separate feature sets for classifying different subcategories of faces by sex. From a computational viewpoint, if there are systematic differences in the information available for sex classification of adults' versus children's faces, the separate feature strategy is optimal because it takes this difference into account explicitly. The optimal computational strategy, however, does not necessarily underlie human performance. It has been suggested that the well-known "other-race effect" for face recognition is the result of misapplying features useful for describing own-race faces to the representation of other-races faces (Malpass & Kravits, 1969; Shepherd, Davies & Ellis, 1981; O'Toole, Deffenbacher, Valentin & Abdi, 1994). Similar arguments have been applied to native versus non-native language learning (Kuhl, Andruski & Chistovich, 1997; Werker & Tees, 1984a,b).

In summary, in the adult feature strategy, the sex classification task is learned from adults' faces. In the child feature strategy, the task is learned from children's faces. In the combined feature strategy, the task is learned from a combination of adults' and children's faces. Finally, in the separate feature strategy, the task is learned separately for adults' and children's faces.

In the present study, we first present simulations of the four strategies for comparison with the pattern of results found in Wild et al. (2000). Two of these strategies yielded results consistent with human performance. Next, to determine which of the two consistent strategy models was a better predictor of human performance, we evaluated model and human performance at the level of individual faces. An experiment with human participants was conducted to obtain classification accuracy for the individual children's faces used in the simulations. Performance on individual faces was then used to assess the accord between each strategy and the performance of human observers.

2. Methods

2.1. Stimuli

Two sets of face stimuli were used to train and test the computational models. The first set of faces included 100 full-face images (50 Caucasian adults and 50 Caucasian children)

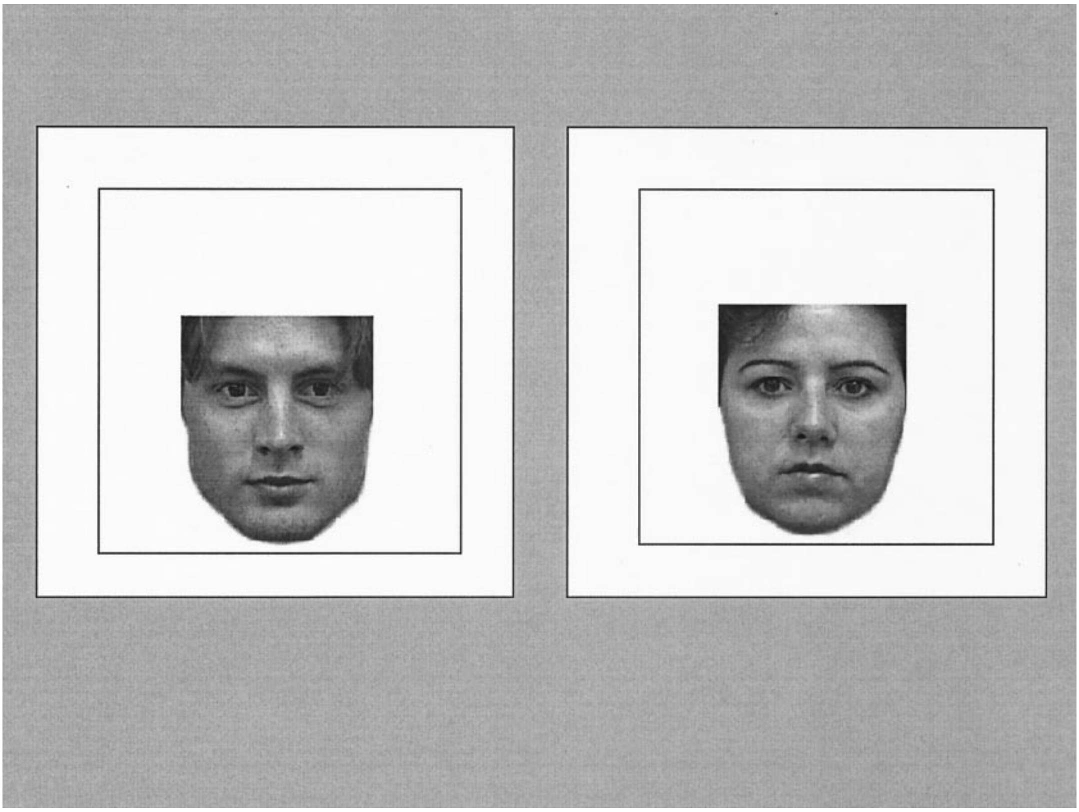


Fig. 2. Sample male (left) and female (right) faces edited to exclude hair and clothing cues.

that were edited digitally to exclude the neck and clothing. Both the adults' and children's face sets comprised equal numbers of male and female faces. The adults' faces were of people in their twenties. The children's faces were of boys and girls between the ages of seven and ten years old. A second face set was created from the same 100 faces, by cropping each face to exclude most of the hair. All pictures were 256 × 256 pixel digital images with a resolution of 256 gray levels per pixel. Examples of the stimuli are displayed in Fig. 2.[2]

2.2. Procedure

The computational procedures used in the present study have been applied in several other studies on adults' faces, with complete details presented in these other papers (Abdi, Valentin, Edelman & O'Toole, 1995; O'Toole, Vetter, Volz & Salter, 1997). In the present study, we give only an overview of the procedures in the body of the paper. However, for completeness, we have included an appendix with details and equations for each of the computational simulations.

All simulations employed the same general method. Each involved: 1) using principal component analysis (PCA) to extract a face representation; 2) applying this representation to

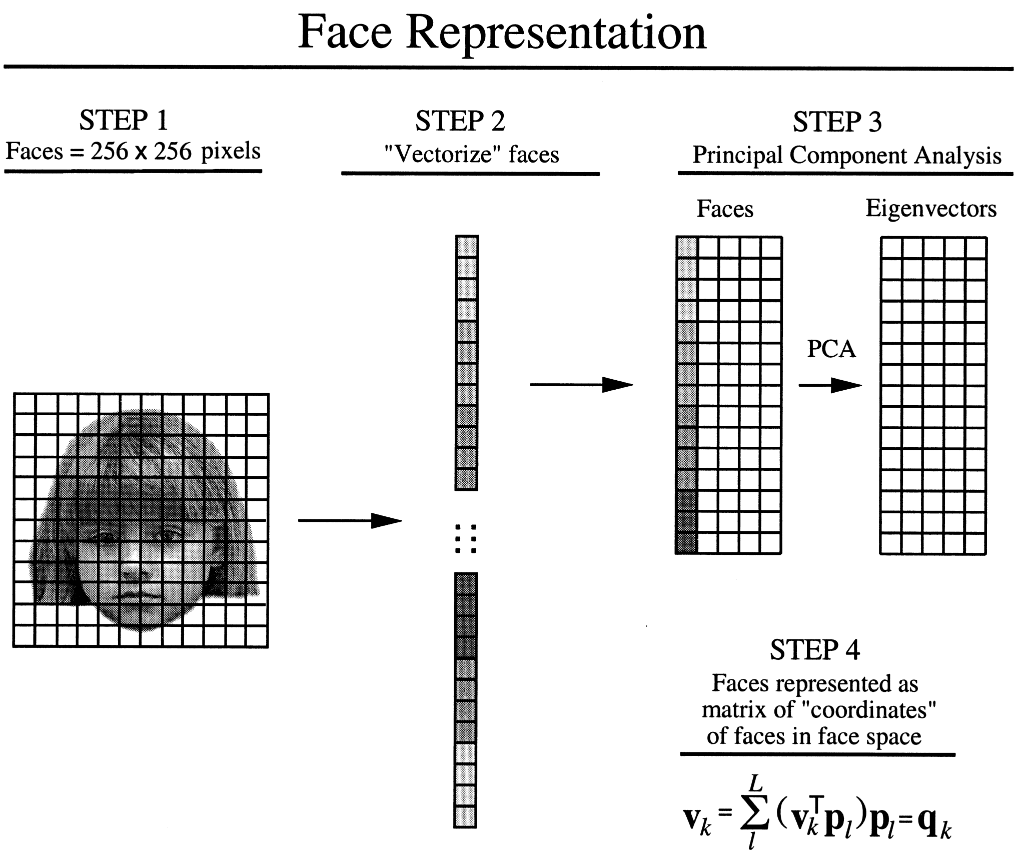
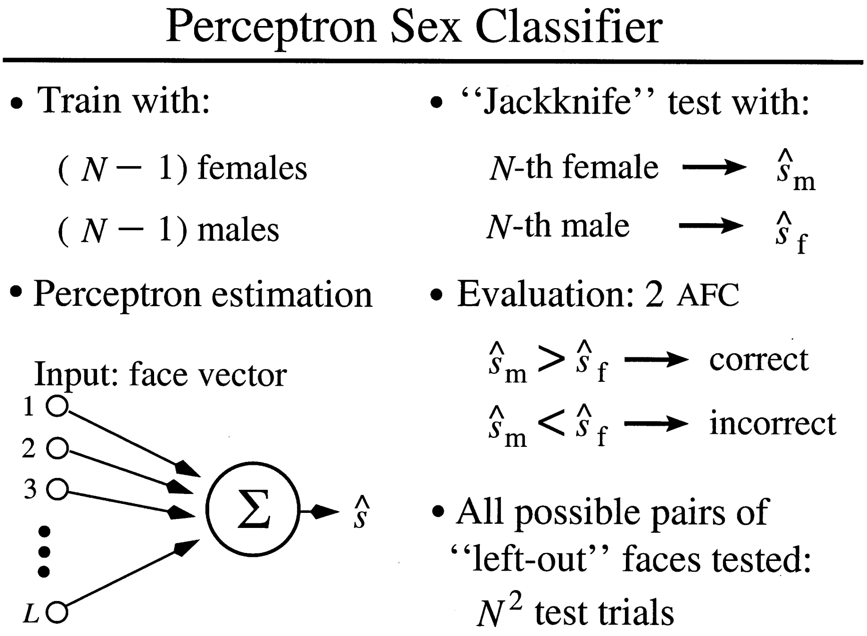


Fig. 3. Schematic presentation of the PCA on faces.

training a perceptron to classify faces by sex; and 3) testing the classification accuracy of the perceptron with a 2-alternative forced choice procedure.

2.2.1. Principal component analysis approach

To extract a face representation, we applied PCA to a set of face images. This process is illustrated schematically in Fig. 3. This yields a set of orthogonal principal components (PC's) or eigenvectors with which the set of faces can be described. The PC's are derived from the statistical structure of the faces and can be ordered according to the proportion of variance they explain in the data. Individual faces from the set can be reconstructed exactly as a weighted combination of the PC's. As such, PC's have been considered analogous to "features" (Abdi, 1988; O'Toole, Abdi, Deffenbacher & Valentin, 1993; Turk & Pentland, 1991) and the weights or coordinates needed to reconstruct a face have been considered feature values. From a geometric point of view, these "features" define the dimensions of a multidimensional "face space" (see also Valentine, 1991), in which individual faces are represented as points in this space. Their position in the space is defined by the coordinates (weights) of their projections onto the eigenvectors. We will refer to this representation



Perceptron Sex Classifier

- Train with:

  $(N-1)$ females

  $(N-1)$ males

- Perceptron estimation

- ‘‘Jackknife’’ test with:

  N-th female ⟶ $\hat{s}_m$

  N-th male ⟶ $\hat{s}_f$

- Evaluation: 2 AFC

  $\hat{s}_m > \hat{s}_f$ ⟶ correct

  $\hat{s}_m < \hat{s}_f$ ⟶ incorrect

- All possible pairs of ‘‘left-out’’ faces tested: N^2 test trials

Fig. 4. Schematic presentation of the perceptron sex classifier.

henceforth as a “face coordinate vector.” In summary, the PCA is used to derive a representation of faces that is dependent on the statistical structure of the faces to which it is applied. In this study, one hundred face images were used to derive the PCs, and consequently, the face coordinate vectors in the combined feature simulation. Fifty face images were used to derive the PCs and the face coordinate vectors for all other simulations.

2.2.2. Linear perceptron classifier

Face coordinate vectors derived from the PCA were used to train two-layer linear perceptrons to classify faces by sex (-1 = female, $+1$ = male).[3] This process is illustrated schematically in Fig. 4. In all cases, the data reported for the sex classification task are from faces that were not used to train the perceptron. In other words, the data represent the ability of the perceptron to *generalize* the sex information learned from a set of faces to classify a (some) novel face(s). For all simulations, the performance of the perceptron model was evaluated with a two-alternative forced choice (2AFC) test using all possible pairs of 25 male and 25 female faces (625 trials). More specifically, for each trial, two novel face coordinate vectors, one male and one female, were input to the trained perceptron and the output activations were computed. A correct response was recorded when the activation for the male face exceeded the activation for the female face. Each data point on the graphs, therefore, represents the proportion of correct responses computed across 625 test trials.

When the perceptron was trained and tested with face coordinate vectors from a single age group, (e.g., training and testing with adults’ faces), the generalization performance of the perceptron was assessed using a “leave-two-out” cross-validation procedure (also called “jackknife”). This procedure was used to make sure that the largest possible sample of novel faces was available for test. The cross validation technique was applied in the following manner. The perceptron was trained on 48 faces (24 male and 24 female faces) out of a total of 50. The perceptron was then tested with the two left-out faces (one male and one female) using the 2AFC procedure described previously. All possible pairs of left-out male and female faces were tested systematically, for a total of 625 trials. When the perceptron was trained and tested with face coordinate vectors of different age faces, (e.g., training with adults' faces and testing with children's faces), the cross validation procedure was not necessary because none of the test faces were in the training set.

Similar to O'Toole et al. (1993) and Valentin, Abdi and Edelman (1994a), we evaluated sex classification performance as a cumulative function of subspace dimensionality by varying the number of eigenvectors used to represent the faces. This was done by training a different set of perceptrons for each subspace. By systematically varying the subspace dimensionality, a more complete picture of the performance of the model is available. An important difference between the present simulations and past work with adults' faces was that we trained all models using faces with hair, but tested using faces without the hair. We did this because we think it best approximates the circumstances in which humans learn the task of sex classification and the circumstances under which psychologists generally test performance. Thus, we typically encounter people in the real world with hair styles that are common for their sex. In comparing our data to the psychological literature on sex categorization of faces, including Wild et al. (2000), faces have been edited to eliminate hairstyle as a cue to sex.

Before proceeding, we note that although the number of faces available may seem small, it has been demonstrated previously (Valentin, Abdi & O'Toole, 1994b; Valentin, Abdi, Edelman & O'Toole, 1997) that the gender of faces can be estimated robustly for adult faces with a similar technique using as few as 10 eigenvectors. These eigenvectors, in turn, are reliably estimated from a similarly small number of faces (Abdi, Valentin & O'Toole, 1997). This is understandable in terms of the large eigenvalues (i.e., proportion of explained variance) associated with gender information in PCA of faces (e.g., O'Toole, Abdi, Deffenbacher & Valentin, 1993). A related technical point concerns the problem of overfitting with the perceptron. This problem is eliminated in the present study by using a sex classification generalization task (i.e., a jackknife). This guarantees that the level of performance cannot be attributed to overfitting.

2.3. Simulations

We implemented the four feature strategies by varying the training sets and test sets. First, PCA was applied to a set of faces. Second, the face coordinate vectors of these faces were computed in the PCA space. Third, a subset of these coordinate vectors was used to train a perceptron sex classifier. Fourth, the perceptron model was tested with coordinate vectors from novel faces using the 2AFC task. Finally, each simulation was performed once for each cumulative subspace dimensionality, incrementally increasing the number of eigenvectors used in representing the faces.

2.3.1. Simulation 1: adult feature strategy

We applied PCA to the adult face set. The face coordinate vectors of the adults' faces were then computed and used to train a perceptron sex classifier. We then tested the perceptron with novel adults' and childrens' faces using the 2AFC procedure. The proportion correct for the adults' versus children's test faces, as a function of subspace dimensionality, is displayed in Fig. 5a. The perceptron's performance for the adults' and children's faces stabilizes at approximately 90% and 70% correct, respectively. The overall level of performance for the model is good and compares favorably to most other models in the literature (for reviews see Abdi, Valentin, Edelman & O'Toole, 1995; O'Toole, Vetter, Volz & Salter, 1997). This is especially true given the two added constraints of transferring learning from faces with hair to faces without hair, and transferring learning between age groups. To our knowledge, this is the first computational model which tests sex classification of faces across a "hair to no hair" transfer condition. For present purposes, the most important aspects of the data can be seen in Fig. 5a. Here it is clear that, with the exception of the first subspace dimensionality, the accuracy for adults' faces is consistently higher than for children's faces as was the case for human subjects (Wild et al., 2000).

2.3.2. Simulation 2: child feature strategy

This simulation was implemented using the procedure described for Simulation 1, except that in this case children's faces served as the training set. The results of Simulation 2 are displayed in Fig. 5b. The figure indicates that classification accuracy for adults' and children's faces peaks at approximately 65% and 85% correct, respectively. In summary, the perceptron classifies children's faces more accurately than adults' faces and is therefore inconsistent with the human behavioral data.

2.3.3. Simulation 3: combined feature strategy

We applied PCA to the combined set of adults' and children's faces (total of 100 faces) and used the resultant face coordinate vectors to train the perceptron sex classifier. Again, the perceptron was tested with novel adults' and children's faces. The results are displayed in Fig. 5c. For subspace dimensionalities up to about ten eigenvectors, the perceptron classified adults' and children's faces at roughly equivalent levels of accuracy. After the first ten eigenvectors, children's faces were classified more accurately than adults' faces. Accuracy peaked at approximately 68% and 82% correct for the adults' and children's faces, respectively. The model is, therefore, inconsistent with the human behavioral data.

2.3.4. Simulation 4: separate feature strategy

We applied two separate PCA's to the adults' and children's faces. The resultant face coordinate vectors were used to train two separate perceptron sex classifiers, one using the adult feature set and the other using the child feature set. In essence, this simulation is made of the adult face test condition from Simulation 1 (see Fig. 5a) and the children's face test condition from Simulation 2 (see Fig. 5b). These conditions appear replotted for comparison in Fig. 5d. Although performance on the first few eigenvectors was generally better for children's faces than adults' faces, once performance stabilized, adults' faces were classified more accurately than children's faces. At its peak, the perceptron classified adults' and

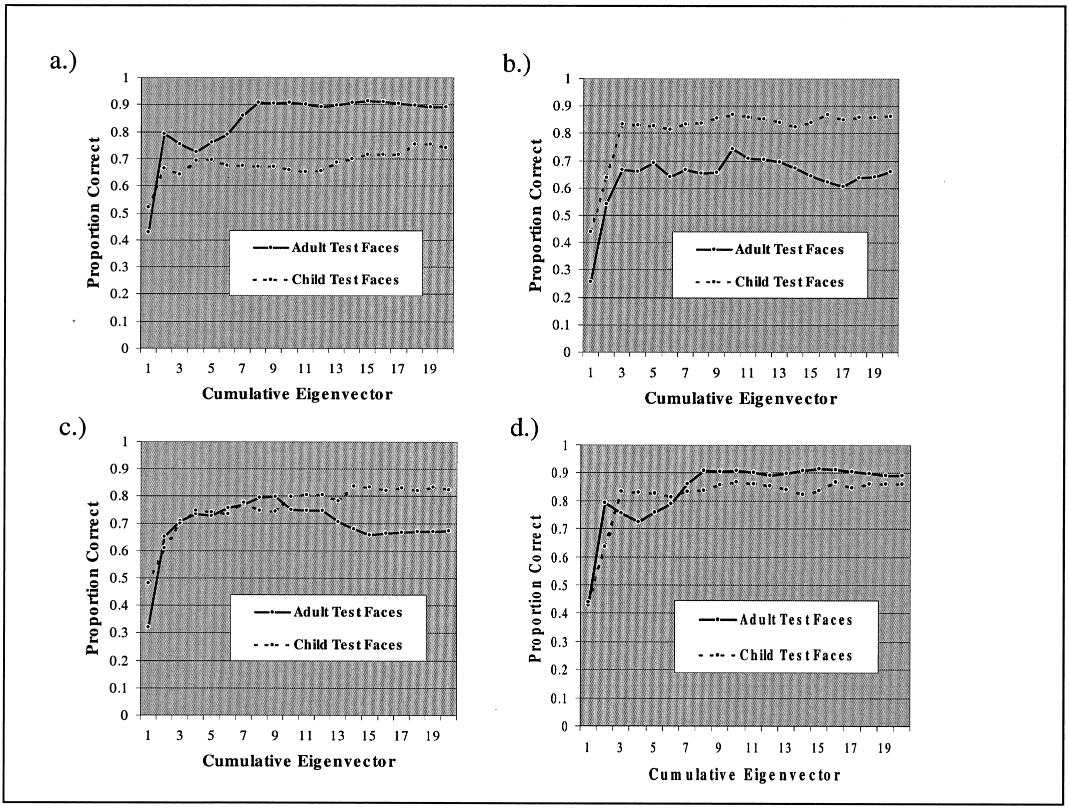


Fig. 5. Simulation results for a.) the adult-based feature strategy; b.) the child-based feature strategy; c.) the combined feature strategy; and d.) the separate feature strategy. The adult face advantage appears for the adult-based and separate strategies.

children's faces at approximately 90% and 85% correct, respectively. Overall, as expected, the separate feature model produced the most accurate performance, Additionally, adults' faces were classified by sex more accurately than children's faces, and so the separate feature strategy is consistent with the performance of human subjects.

2.3.5. Simulation discussion

From the simulations we can make some general statements about the relationship between the sex information in adults' and children's faces. First, given that the performance of the simulations was well above chance in all of the face age transfer conditions, we can eliminate the extreme hypothesis that there is no shared sex information in adults' and children's faces. Second, at the other extreme, there was always a cost associated with transferring sex classification between learning and testing conditions across age groups in the simulations. Thus, we can rule out the somewhat trivial possibility that the sex information for adults' and children's faces is equivalent.

Before discussing the models that were consistent with the psychological data, some useful points can be made from the two models that were eliminated. One might argue that the task of learning to classify faces by sex is acquired during childhood and thus might be strongly influenced by the information available in the faces of other children. However, the child feature strategy did not reproduce the advantage for adults' faces. In addition, the results of Wild et al. (2000) indicate that even young children show this advantage for adult's faces. Thus, we can conclude that children's faces do not form the basis of our learned knowledge about the sex of faces.

When there is consistent information for making categorical decisions in overlapping categories, a parsimonious approach to the problem would be to combine this information in an optimal way to operate across categories. The performance of the combined feature model on adults' and children's faces in low subspace dimensionalities was roughly equivalent. From a performance point of view, this finding suggests that a compromise encoding of adults' and children's faces is indeed possible. However, beyond low subspace dimensionalities, children's faces were classified more accurately than adults' faces. Thus, the model performance is inconsistent with the experimental results for all of the subspace dimensionalities tested. The combined feature simulation indicates that forcing the model to settle on a feature set that optimally classifies both the children's and adults' faces does not reproduce the classification advantage for adults' faces seen in the psychological data. It is perhaps worth noting, also, that overall the combined model did not achieve better performance than the other models. This suggests that in settling on an encoding suitable for both children's and adults' faces, some features useful for only one age group of faces may be applied inappropriately to the other age group of faces. This can happen in the combined model because of the importance of explained variance, *relative* to the statistical structure of the training set, in determining the feature set (e.g., eigenvectors) derived from the PCA. When faces from both age groups comprise the training set, and when the importance of individual features for the two age groups is not equivalent, the model will converge on a compromise that may over- or underweight these features for one or both age groups in the classification decision.

The adult feature strategy produced data consistent with the general pattern of performance found in human subjects. This strategy instantiates the hypothesis that there is shared information for classifying adults' and children's faces by sex and that a useful subset of this information can be extracted from adults' faces. This strategy produced above chance performance for both adults' and children's faces and performed more accurately for adults' faces. Thus, it is possible that adults' faces could form the basis of the feature set we apply to classifying the sex of children's faces.

The separate feature strategy represents a rather different approach to the problem, but also produced data that were generally consistent with human performance. While allowing for the possibility that there is shared sex information for adults' and children's faces, the separate feature model does not make explicit use of this information, Thus, this strategy is tailored to serve the perceptually divergent needs of conceptually unified categorizations that must be carried out in different subcategories. Simply put, although the gender concept may be a consistent and unified idea applicable to faces of all ages, the perceptual problem of determining gender from faces within different age categories may be solved in a less unified way.

3. Human experiment

Human subjects from Wild et al. (2000) classified adults' faces more accurately than children's faces. The adult and separate feature models were consistent with these data. The next logical step was to determine which of these strategies is a better model of human performance. To do this, we evaluated the performance of these models at the level of individual faces. This enabled us to compare the pattern of errors that the models and human subjects make. In other words, are the model and humans erring on the same faces or on different faces? This enabled us to distinguish between the remaining two models that are in general agreement with the human data.

We conducted an experiment with human subjects classifying children's faces by sex. Only children's faces were used, because the data of Wild et al. (2000) indicated nearly perfect accuracy for classifying the adults' faces. This obviously limits the ability of adults' faces to provide individual face predictions that will help us decide between the two models. Note that the purpose of this experiment was to gather data on the accuracy with which *individual faces* were classified by sex.

3.1. Participants

Thirteen undergraduates (6 female and 7 male) from the University of Texas at Dallas volunteered to participate in the experiment. All subjects received one experimental credit required by a core course in the psychology program.

3.2. Stimuli

The stimuli consisted of the 50 children's faces used in the simulations as test faces. Note that these were faces without hair, (see Simulation Methods for details).

3.3. Procedure

Participants were asked to classify all 50 faces by sex. The faces were presented one at a time on a computer screen. Each face remained on the screen until the subject pressed one of the labeled keys (male or female) on the computer keyboard. The order of face presentation was randomized for each observer. All experimental events were controlled by a Macintosh computer programmed with Psyscope (Cohen, McWhinney, Flatt & Provost, 1993).

4. Human experiment results

The results were analyzed in two ways. First, for comparison with the data of Wild et al. (2000), we analyzed the accuracy of the human participants, collapsing across the individual faces. Second, and more directly relevant for the present study, we analyzed accuracy for individual faces, collapsing across the individual participants.

4.1. Participant analysis. Sex classification accuracy was measured for each observer using A' for discriminating male and female faces. The statistic A' (a nonparametric version of the signal detection measure of d′) was appropriate for two reasons. First, both Wild et al. (2000) and Intons-Peterson (1988) found a consistent response bias for guessing "male." We will consider this bias in more detail in the next section. For present purposes, a bias-free measure was needed to get a clear picture of discrimination accuracy. Second, A' was used by Wild et al. because of the relatively high accuracy they had for the classification of adults' faces.[4] Although not as much of a problem with the children's faces, we used A' for comparability with that study.

The A' score was calculated based on "hits," defined arbitrarily as the response "female" to female faces, and "false alarms" defined as the response "female" to male faces. Overall accuracy for classifying the children's faces by sex was reasonably good, $A' = 0.85$, and roughly comparable to the analogous condition in Wild et al. (2000), $A' = 0.77$. Slight differences in performance may be due to the use of different face sets.

We also found a bias for responding "male," as evidenced by the average criterion across individual participants, which was significantly greater than zero, $C = 0.43$, $t = 6.88$, $p <$ .001. A positive C score indicates a bias to respond "male" and a negative score indicates a bias to respond "female." This bias was quite consistent with all but one of the participants having a positive C value. Overall, 61.9% of the responses were "male." The sex of the participant was not related to the magnitude of the response bias, $F(1,11) < 1$, $MS_e = 0.55$.

4.2. Face analysis. The face analysis proceeded as follows. The average percentage correct was calculated for each face collapsing across individual participants. Note that our design does not allow the computation of a bias-free measure (such as A') or criterion (such as C) for individual faces because each face has a single determined sex and thus cannot serve both as signal (e.g., female) and noise (e.g., male). However, a response bias is still present in the data and manifests itself in the generally better performance for male (88.6% correct) versus female (64.5% correct) faces. This difference was significant, $F(1,48) = 12.89$, $MS_e = 0.649$, $p < .001$. We note this lack of a bias-free measure for individual faces because it has implications for how we implemented the human-model individual face comparison.

5. Model performance on individual faces

Given a measure of human accuracy on each face, the next step was to extract a measure of model classification accuracy for each face. Recall that the simulation models were tested with a 2AFC task in which a novel male and female face were input to the perceptron and the activations of the output unit were compared. The classification decision was considered correct if the activation for the male face exceeded the activation for the female face. This yielded a percentage correct across faces that was based on the bias-free 2AFC task. Although this measure is comparable to the bias-free A' computed by Wild et al. (2000) for human observers, it is not a measure that can be used with individual faces. We proceeded, therefore, to formulate a measure that was more analogous to the human task of deciding if an individual face is male or female. We describe this method in general terms first, and then apply it to the individual model comparisons.

Model accuracy for the individual children's faces was measured by computing the perceptron activations for each novel test face and comparing these activations to a threshold or criterion activation. Faces with activations above the criterion were classified as male, whereas faces with activations equal to or below the criterion were classified as female. The criterion activation was set at a level comparable to the male response bias found in the human experiment. This was done by noting the average activation for male faces and for female faces and by bisecting this to achieve the ideal observer criterion.[5] From here, the model criterion was shifted in the male bias direction until the overall performance rates matched those found with the human participants for male and female faces.

More specifically, accuracy for individual children's faces was tested for the adult feature strategy and the separate strategy for all 50 faces, presented as novel to their respective perceptrons. The subspace dimensionality tested for the perceptrons was set to 18 eigenvectors, the point at which the model performance was stable for all four simulations.[6] This yielded a correctness value (1 or 0) for each face, for each feature strategy.

6. Model versus human results

Our goal was to compare the pattern of errors for the two feature strategies to the pattern of errors made by human subjects. Separate ANOVA's were carried out on the adult and separate feature strategy data using the model correctness as a predictor of the human performance on the individual faces. The results indicated that the separate feature strategy proved a reliable predictor of human accuracy, $F(1,48) = 4.68$, $MS_e = 0.057$, $p < .05$, whereas the adult feature strategy did not $F(1,48) < 1$, $MS_e = 0.063$. To get a better perspective on the strength of the model-human relationship for the two strategies, we compared the model-human correlation for accuracy on individual faces to the average correlation between the individual subjects for accuracy on individual faces. For the separate feature strategy, we obtained a correlation coefficient of $r = 0.30$, $p < .05$, between the model and human performance on the individual faces. To compare this correlation to the consistency of individual participants, we computed the average correlation among error patterns for all possible pairs of participants. The result was an average correlation of $M_r = .29$, with a standard deviation of .02. Thus, the correlation between the separate feature strategy and human performance, though moderate in size, was about equal to the correlation between individual participants. By comparison, for the adult feature strategy, we obtained no correlation between the model and human performance on the individual faces, $r = .01$.

Finally, it is perhaps worth noting the difficulty of discriminating boys' and girls' faces without the presence of sex-stereotyped cues. This is best illustrated by considering human percentage correct scores for the individual faces. In fact, only about 20% of the 50 faces were always classified correctly. At the other extreme, between 5 to 10% of the faces were consistently misclassified. In all cases, these were girls who were misclassified as boys. In between we found accuracy values that varied across the spectrum. These data illustrate that classifying children's faces without sex-stereotyped cues is a difficult task that varies considerably as a function of individual faces.

7. General discussion

The first purpose of this study was to quantify and compare the information available in adults' and children's faces for sex classification. The simulations eliminated the two extreme hypotheses of complete overlap and no overlap in the sex information in adults' and children's faces, indicating that there is shared information for sex categorization for these age groups. This evidence can be found in the results of the adult and child feature simulations, both of which showed above-chance sex classification generalization to the face age category that was not learned. The best overall performance for children's and adults' faces, however, was achieved in the separate feature strategy, in which the features were tailored to the appropriate age group of faces. The superior performance of the model on adults' faces in this case indicates that the information for sex classification in adults' faces is inherently more reliable than the information in children's faces.

The second purpose of the study was to evaluate the four feature strategies as psychological models of human performance on the task of classifying adults' and children's faces by sex. The data of Wild et al. (2000) indicate that sex classification accuracy was better for adults' faces than for children's faces. In evaluating the simulations as psychological models of sex classification, although all four models performed the tasks at levels above chance, only the adult feature strategy and the separate strategy reproduced the advantage for adults' faces found in human subjects (Wild et al., 2000). The child feature strategy and the combined models were inconsistent with the psychological data.

The two models that were consistent with the psychological data, (i.e., the adult feature and the separate feature models), make use of rather different feature generalization strategies. The adult feature model generalizes features nonoptimally, but can operate with a single set of features. As noted previously, the other-race effect has been hypothesized to be caused by the application of a statistically inappropriate feature set for encoding faces. The simulations indicate that generalization from adults' to children's faces works reasonably well and produces data that are consistent with psychological data showing a classification advantage for adults' faces. The separate feature model does not require generalization, but employs two sets of noninteracting features. The superior performance of this model on adults' faces can be accounted for by the inherent reliability of the sex information in adults' versus children's faces, without recourse to more complicated processing theories for making use of the shared information.

The general agreement between the psychological data and these two rather different feature generalization strategies led us to look more closely at human performance on the task. Using the pattern of errors on individual faces, we found that the separate feature strategy provided a better fit to human data. Thus, the models suggest that human performance is best fit by a processing strategy that is optimized separately within subcategories. The separate model's accord with the human data at the level of individual faces suggests that human errors for classifying faces by sex do not have their source in the misapplication of age inappropriate sex-related features. Although less parsimonious than the other three models tested, which achieve classification with a single unified set of features, this model is better in terms of overall performance and better predicts human classification errors on individual faces. The separate model also suggests that sex and age classification may be linked in an interesting way, implying that age classification may precede sex classification.

In the broader context, faces constitute an important and highly meaningful category of objects. This category is further subdivided into multiple and overlapping subcategories based on the sex, race, and age of faces. These are socially meaningful categories that one detects through perceptual processes. Although sex, race, and age share similarities as visually-derived semantic "categories" of faces (Bruce and Young, 1986), the information that specifies a face's status with respect to these dimensions interacts dynamically over time. As we age, the quality, the quantity, and the form of the gender-specifying information in our face changes. The markers of age are also dynamic and may be categorical for some age distinctions and more continuous for other distinctions. Understanding the complexities of the perceptual information that specifies category membership, and the relationship among multiple coexisting categories, is a prerequisite to sorting through plausible theories of the ways in which humans perform these tasks.

Notes

1. The first-graders were between 6 and 8 years old, and the third-graders were between 8 and 10 years old.
2. We included only example stimuli from adults' faces because permission was not requested from parents to publish the children's faces. However, the children's faces were processed and presented in exactly the same manner as adult faces.
3. The perceptron is a simple linear deterministic neural network that reaches a unique solution (Abdi, Valentin & Edelman, 1999). Strictly speaking, we have used a linear hetero-associator, but the term perceptron is generally used when the predicted values are binary (see Abdi, 1994).
4. At high levels of performance, the nonlinearities of d' can give a distorted view of difference between conditions.
5. This corresponds to the ideal observer as defined by signal detection theory (cf. Green & Swets, 1966).
6. We also explored neighboring subspace dimensionalities with similar results. Note that beyond 20–25 eigenvectors, overfitting became a factor for some of the perceptrons.

Acknowledgments

Thanks are to the National Institutes of Standards and Technology for providing support to A.J. O'Toole and Y.D. Cheng during the completion of this project.

Appendix

In this appendix we give a formal description of the techniques used in the simulations.

We start with a set of $K = 2N$ digitized face images with half of the faces being male and half being female. Each face image is represented by the column vector of the I pixel light intensities (e.g., a 256^2 vector). The set of face images is therefore an $I \times K$ matrix denoted $\mathbf{X}$.

The first step is to obtain the singular value decomposition of $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^\top \qquad \text{with: } \mathbf{P}^\top\mathbf{P} = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I} \tag{1}$$

where $\mathbf{P}$ (respectively $\mathbf{Q}$) is the matrix of the left (respectively right) singular vectors of $\mathbf{X}$, and $\boldsymbol{\Delta}$ is the diagonal matrix of the singular values of $\mathbf{X}$. Because both $\mathbf{P}$ and $\mathbf{Q}$ can be obtained using the technique of eigendecomposition, they are frequently called eigenvector matrices in the PCA literature. The matrix $\mathbf{Q}$ gives also the (normalized) projections of the faces on $\mathbf{P}$ (because, $\mathbf{Q} = \mathbf{X}^\top\mathbf{P}\boldsymbol{\Delta}^{-1}$, see for more details, e.g., Abdi, 1988; Abdi, Valentin & Edelman, 1999).

For a given simulation we set the number of projections (i.e., the number of eigenvectors) that we want to keep. Call this number L. The *face coordinate vectors* will correspond to the first L columns of $\mathbf{Q}$. The $K \times L$ matrix storing the face coordinate vectors is called $\mathbf{V}$. It is formally defined as

$$\mathbf{V} = [v_{k,l}] = [q_{k,l}] \qquad \text{for } k = \{1 \cdots K\}, \text{ and } l = \{1 \cdots L\}. \tag{2}$$

The learning set of faces is composed of $K - 2$ faces (the original K faces minus one male face and minus one female face). The 2 faces not present in the learning set are used to test the performance of the model. Assume that the male (respectively female) face is the m-th face (respectively f-th face), and that the corresponding row face vectors are denoted $\mathbf{m}^\top$ and $\mathbf{f}^\top$. The set of the learned face coordinate vectors is stored in a matrix denoted $\mathbf{V}_{\text{learn}}$ and the 2 face coordinate vectors for the test faces are stored in a matrix denoted $\mathbf{V}_{\text{test}}$. Formally, this is equivalent to partitioning the matrix $\mathbf{V}$ as:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{\text{learn}} \\ \mathbf{V}_{\text{test}} \end{bmatrix}, \qquad \text{with: } \mathbf{V}_{\text{test}} = \begin{bmatrix} \mathbf{m}^\top \\ \mathbf{f}^\top \end{bmatrix}. \tag{3}$$

The perceptron (or discriminant analysis) learning technique is implemented by associating to the learn faces a dummy sex vector $\mathbf{s}$. The value of the elements of $\mathbf{s}$ is $+1$ for males and -1 for females. Learning is equivalent to predicting $\mathbf{s}$ from $\mathbf{V}_{\text{learn}}$. This boils down to finding a vector of weights, denoted $\mathbf{w}$ such that

$$\hat{\mathbf{s}} = \mathbf{V}_{\text{learn}}\mathbf{w}^\top \qquad \text{with } (\hat{\mathbf{s}} - \mathbf{s})^\top(\hat{\mathbf{s}} - \mathbf{s}) \text{ being minimum.} \tag{4}$$

This is obtained as

$$\mathbf{w} = \mathbf{V}_{\text{learn}}^{+}\mathbf{s} = \mathbf{V}_{\text{learn}}\mathbf{s} \text{ (because } \mathbf{V}^\top\mathbf{V} = \mathbf{I}, \text{ and therefore } \mathbf{V}^{+} = \mathbf{V}^\top), \tag{5}$$

(with $\mathbf{V}_{\text{learn}}^{+}$ being the More-Penrose pseudoinverse of $\mathbf{V}_{\text{learn}}$). The next step is to compute the predicted values (or activation in a neural network framework) of the dummy sex indicator as

$$\hat{\mathbf{s}}_{\text{test}} = \begin{bmatrix} \hat{s}_{\text{male}} \\ \hat{s}_{\text{female}} \end{bmatrix} = \begin{bmatrix} \mathbf{m}^{\top}\mathbf{w} \\ \mathbf{f}^{\top}\mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{m}^{\top} \\ \mathbf{f}^{\top} \end{bmatrix}\mathbf{w} = \mathbf{V}_{\text{test}}\mathbf{w}. \tag{6}$$

We then use the values of $\hat{\mathbf{s}}_{\text{test}}$ to predict the sex of the test faces with the following rule: If $\hat{s}_{\text{male}} \geq \hat{s}_{\text{female}}$, then the male face is identified as a male and the female as a female. This corresponds to a correct classification. If, on the contrary, $\hat{s}_{\text{male}} < \hat{s}_{\text{female}}$, then the male face is identified as a female and the female as a male. This corresponds to a classification error.

References

Abdi H. (1988). A generalized approach for connectionist auto-associative memories: interpretation, implication and illustration for face processing. In J. Demongeot T. Hervé V. Rialle C. Roche (Eds.) *Artificial intelligence and cognitive sciences* (pp. 149–166). Manchester, UK: Manchester University Press.

Abdi H. (1994). *Les réseaux de neurones.* Grenoble, France: Presses Universitaires de Grenoble.

Abdi H., Valentin D., & Edelman B. (1999). *Neural networks.* Thousand Oaks, CA: Sage.

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

Abdi H., Valentin D., & O'Toole A. (1997). A generalized autoassociator model for face processing and sex categorization: from principal components to multivariate analysis. In D. Levine W. R. Elsberry (Eds.) *Optimality in biological and artificial networks* (pp. 317–337). Hillsdale, NJ: Lawrence Erlbaum.

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

Cohen J. D., McWhinney B., Flatt M., & Provost J. (1993). Psyscope: a new-graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments and Computers, 25,* 257–271.

Cornell E. H. (1974). Infants' discrimination of photographs of faces following redundant presentation. *Journal of Experimental Child Psychology, 18,* 98–106.

Enlow D. (1982). *Handbook of facial growth.* Philadelphia, PA: Saunders.

Fagot B., & Leinbach M. (1993). Gender-role development in young children: from discrimination to labeling. *Developmental Review, 13,* 205–224.

Green D. M., & Swets J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Intons-Peterson M. (1988). *Children's concepts of gender.* Norwood, NJ: Ablex.

Kuhl P. K., Andruski J., & Chistovich I. A. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277,* 684–686.

Malpass R., & Kravitz J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology, 13,* 330–334.

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

O'Toole A. J., Deffenbacher K. A., Valentin D., & Abdi H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition, 22,* 208–224.

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., Volz H., & Salter E. M. (1997). Three-dimensional caricatures of human heads: distinctiveness and the perception of facial age. *Perception, 26,* 719–732.

Shepherd J., Davies G., & Ellis H. (1981). Studies of cue saliency. In G. Davies H. Ellis J. Shepherd (Eds.) *Perceiving and remembering faces* (pp. 105–131). London: Academic.

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

Valentin D., Abdi H., & Edelman B. (1994). What represents a face: a computational approach for the integration of physiological and psychological data. *Perception, 26,* 1271–1288.

Valentin D., Abdi H., Edelman B., & O'Toole A. J. (1997). Principal component and neural network approach: what can be generalized in gender classification. *Journal of Mathematical Psychology, 41,* 398–412.

Valentin D., Abdi H., & O'Toole (1994). Categorization and identification of human face images by neural networks: a review of the linear auto-associator and principal component approaches. *Journal of Biological Systems, 2,* 224–258.

Valentine T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A,* 161–204.

Werker J. F., & Tees R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development, 7,* 49–63.

Werker J. F., & Tees R. C. (1984). Phonemeic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America, 75,* 1866–1878.

Wild H., Barrett S., Spence M., O'Toole A., Chenh Y., & Brooke J. (2000). Recognition and sex classification of adults' and children's faces: examining performance in the absence of sex-stereotyped cues. *Journal of Experimental Child Psychology, 77,* 269–291.

Yamaguchi M. K., Hirukawa T., & Kanazawa S. (1995). Judgement of gender through facial parts. *Perception, 24,* 563–575.