Journal of Biological Systems, Vol.4, No#, pages ***–***. ©

SEX CLASSIFICATION OF FACE AREAS: HOW WELL CAN A LINEAR NEURAL NETWORK PREDICT HUMAN PERFORMANCE?

BETTY EDELMAN, DOMINIQUE VALENTIN†, HERVÉ ABDI†

School of Human Development, The University of Texas at Dallas, Richardson, TX 75083–0688, U.S.A., †also Université de Bourgogne, 21004, Dijon, France.

ABSTRACT

Human subjects and an artificial neural network, composed of an autoassociative memory and a perceptron, gender classified the same 160 frontal face images (80 male and 80 female). All 160 face images were presented under three conditions 1) full face image with the hair cropped 2) top portion only of the Condition 1 image 3) bottom portion only of the Condition 1 image. Predictions from simulations using Condition 1 stimuli for training and testing novel stimuli in Conditions 1, 2, and 3, were compared to human subject performance. Although the network showed a fair ability to generalize learning to new stimuli under the three conditions, performing from 66 to 78 percent correctly on novel faces, and predicted main effects, a more detailed comparison with the human data was not as promising. As expected, human accuracy declined with decreased image area, but showed a surprising interaction between the sex of the face and the partial image conditions. The network failed to predict this interaction, or the likelihood of correct human classification for a particular face. This analysis on an item level raises concern about the psychological relevance of the model.

Keywords: Face, sex classification, neural network.

The authors thank W.K. Estes and J.R. Vokey for suggestions to improve the psychological relevance of model classification, K. Deffenbacher and J.R. Vokey for helpful comments on previous drafts of this paper, and June Chance and Al Goldstein for providing the faces used as stimuli. Portions of this paper were presented at the 1996 Mathematical Psychology meeting in Chapel Hill North Carolina and as a poster at the 1996 Psychonomics conference in Chicago.

Correspondence should be addressed to Betty Edelman, School of Human Development, The University of Texas at Dallas, MS: GR4.1, Richardson, TX 75083–0688, email bedelman@utdallas.edu.

1. Introduction

When a model, such as an artificial neural network, predicts human behavior on average, it is frequently judged successful. However, this success does not necessarily indicate psychological relevance. The model, although capable of producing overall results comparable to human subjects, may be functioning only partially, or not at all, like human beings. A finer grained analysis of model predictions and human performance can serve to examine these possibilities. Such an undertaking, in the domain of sex classification of faces, is the aim of this work.

Recent linear neural network simulations by Abdi, Valentin, Edelman, and O'Toole [3] using photographic image pixel-based input, preprocessed via eigen-decomposition, were able to sex-classify pictures of human faces with 90% accuracy. This result compares favorably to those attained with human subjects by Bruce and her colleagues of 96% accuracy [5], and measurement-based models, such as that of Brunelli and Poggio [7], which was 90% accurate on faces in a training set and 79% accurate on novel faces. However, whereas the face images employed by Bruce et al. [5] had the hair concealed by a swimming cap and the Brunelli and Poggio coding did not include measurement of hair, the stimuli used by Abdi et al. [3] included the hair. To determine the role of hair shape and length in the performance of the Abdi et al. model, further simulations were run using the same face stimuli with the hair masked. Without the hair, the classification accuracy of the model declined to about 80%. However, this performance remained superior to that of an unsupervised classifier (k-means) used as a control measure, which resulted in only 56% accuracy.[1] Therefore, although the hair style is an important contributor to the model performance, the authors concluded that it is not the sole contributor. Other face characteristics contributing to the model performance, and the psychological relevance of the model, remained in question.

The purpose of this work was to investigate further the performance of the Abdi et al. model in regard to the above unresolved issues. First, to assess whether facial areas contribute differently to the model's sex classification performance, we conducted additional simulations, masking first the bottom and then the top of the face area, in addition to the hair. Second, we tested the ability of human subjects to classify the same stimuli as used in these simulations. To determine the psychological relevance of the model, we compared the predictions of the model to the results from human subjects, on both a global and an item level.

Previous simulations by Fleming and Cottrell [8], using masking of bottom and top face areas, with a more computationally demanding nonlinear approach, were not strikingly accurate for sex classification under either masking condition, but showed better performance on the top portion of the face. When the top of the face was masked the model was 29% correct, and when the bottom of the face was masked the model was 55% correct. However, this relatively low performance could be attributed, at least partially, to the great variation of training stimuli, which

[1] The k-means classifier is considered unsupervised because no explicit information about class membership is given to the algorithm, only the number of classes is specified *a priori*.

included non-face images. One possible reason for the difference found between top and bottom conditions is that the Fleming and Cottrell stimuli included the hair, and thus provided more variation in the lower portion of the image, particularly for female faces.

Previous studies of facial area contribution to sex classification by human subjects from photographic images have used several approaches: presenting features (or a combination of features) in isolation [6, 11], masking features [5, 11], and replacing features within a full image [6, 16]. In some cases, the studies have used individual photographic images, and in other cases, male and female prototypes have been created using various averaging techniques. These studies have produced varying results. Differences obtained between tests of features in isolation and substitution of features have been attributed to the role of configuration in facial tasks [6, 11]. For example, although the nose alone provides little information, masking it diminishes the total amount of configural information perceived. In general, these studies indicate that the isolated areas contributing the most to sex classification are: the eye region (particularly the eyebrows), and the face outline (particularly the jaw).

In what follows, we describe the simulations and the predictions obtained from them, followed by a human subject experiment using equivalent stimuli. The discussion of the experimental results compares them to the predictions derived from the model performance, from the highest level of statistical main effects down to the lowest level of classification decisions for individual stimuli.

2. Simulations

These simulations were conducted to investigate the contribution of different facial areas to sex classification by a simple linear model, and to establish predictions for human performance. To approximate the experience human subjects bring to this task, the model was trained using full face images, and then tested on novel full face and partial images.

If, as suggested before, inclusion of the hair in the Fleming and Cottrell [8] simulations contributed to poorer performance when the top of the face was masked, masking the hair in these simulations should yield less difference in accuracy between top and bottom areas. Masking of the hair may also modify the bias toward male classification observed by Abdi et al. [3]: The model, when tested with full face images including the hair, was more accurate in classifying males (about 95%) than females (about 88%). One possible explanation for this result is the larger variation in the female stimuli hair styles. Reducing this male-female distinction, by masking the hair, should greatly reduce this discrepancy for the model, if variation of hair style is indeed the primary cause of better model performance on male faces. On the other hand, if model performance remains superior for male faces, the contribution of other factors to this difference should be investigated.

Most importantly, even if the model shows some ability, under varying conditions, to sex classify the novel stimuli correctly, it may not be a good predictor of human performance. In other words, the classification given to a particular face in a particular condition may have little, or no, power to predict human classification. Therefore, what is of particular interest is the individual face classifications under each stimulus condition.

2.1. ***Model Description***

The model is composed of a linear autoassociative memory and a classifier. The memory, which simulates the storage and retrieval of patterns, can be implemented in several ways. We present a conceptual background for two equivalent approaches: a neural network and principal component analysis. Likewise, the sex classification of a face retrieved from the memory can be simulated using different techniques. We decided to use a simple perceptron to perform this task. For those interested, some additional mathematical details are provided in an appendix.

2.1.1. *Model Input*

The input to the model, for both learning and testing, is digitized from black and white photographic images of faces. Each face image is transformed into a vector (i.e., a column of numbers) by rearranging the picture element (pixel) gray-level values into one column, as illustrated by the left side of Figure 1. The vectors are then normalized so that the length of each face vector is unity. This normalization provides some computational convenience, and also serves to mitigate any differences in overall illumination level.

2.1.2. *Autoassociative Memory - Neural Network Perspective*

In the neural network implementation of an autoassociative memory, the memory is composed of a set of units, which may be compared loosely to neurons, or clusters of neurons. Each unit can take on a numeric value or *activation level.* For this particular autoassociative memory, each unit corresponds to a specific pixel position in a face image, with its activation level representing a gray-level intensity. All units have a weighted connection to all other units, including themselves, as illustrated by the right side of Figure 1. The weights on these connections can be conveniently arranged in a matrix (i.e., a table of numbers) having a row and a column for each unit. For a face image consisting of 156 rows and 99 columns of pixels there are 15444 connected units, and therefore the weight matrix is 15444×15444.

There are two phases in the performance of an autoassociative memory: learning and testing. First, in the learning phase, the memory is presented with a pattern, such as a face image, and develops the connection weights that reflect the relationships between the pattern elements. More than one pattern can be "learned" and the same set of connection weights will reflect the unit relationships for each of the patterns. When the learning phase has established the connection weights, the memory is tested by determining how well it can reconstruct a learned pattern, or a degraded version of a learned pattern, or even a novel pattern.

The testing phase, illustrated in Figure 2, is initiated by setting each memory unit to the value of an element in a memory cue or "key." The activation level of each unit is then calculated by summing the weighted values (i.e., unit value

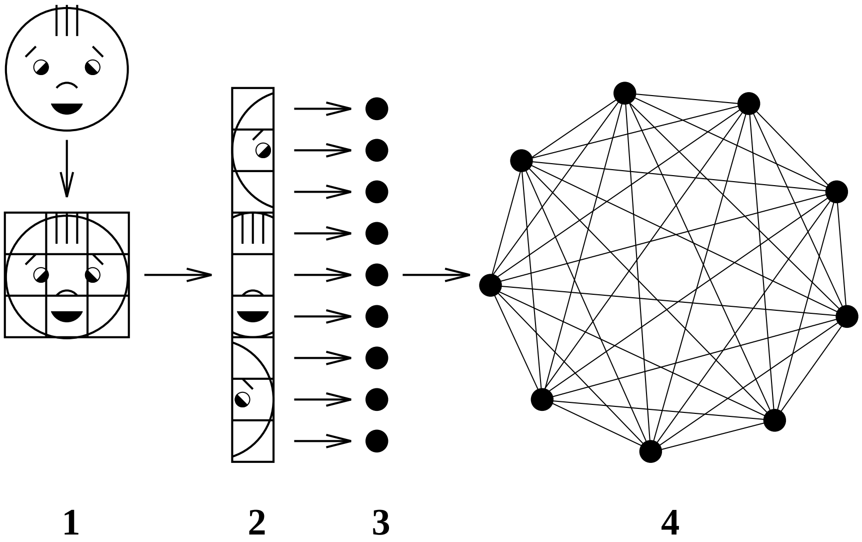

FIGURE 1. Application of an autoassociative memory to face image processing. 1) A face is digitized. For ease of illustration, this face is digitized into only 9 "macro-pixels". 2) The gray-level values of the pixels are rearranged to form a column vector. 3) Each unit of the memory is set to a pixel value. 4) All units are connected to all other units by weight values to form the memory.

times connection weight) of all connected units. The unit activation levels resulting from this integration are interpreted as the response of the memory. The quality of this testing phase reconstruction depends upon the similarity of the memory key to the learned patterns, and also on the technique used to establish the connection weights.

Although there are several learning rules that can be used to establish the connection weights, the most well known are the Hebbian and the Widrow-Hoff algorithms. We will discuss the Hebbian rule first. In this learning algorithm the weight on a connection between two units is simply set to be proportional to the correlation between the values of the units (pixels in this case) in the patterns presented during learning. In other words, the connection weights represent the strength of the relationship between the values of the two units, or the ability of one unit to predict the value of the other.

Although Hebbian learning affords a relatively straightforward interpretation, and computational convenience, it also has its drawbacks. As the number of patterns to be learned increases, the quality of test phase reconstruction may decrease due to interference or "cross-talk" between patterns. This decline of reconstruction quality occurs when the learned patterns contain similarities (they are not pairwise orthogonal or independent of each other; in other words, their cross-product is not

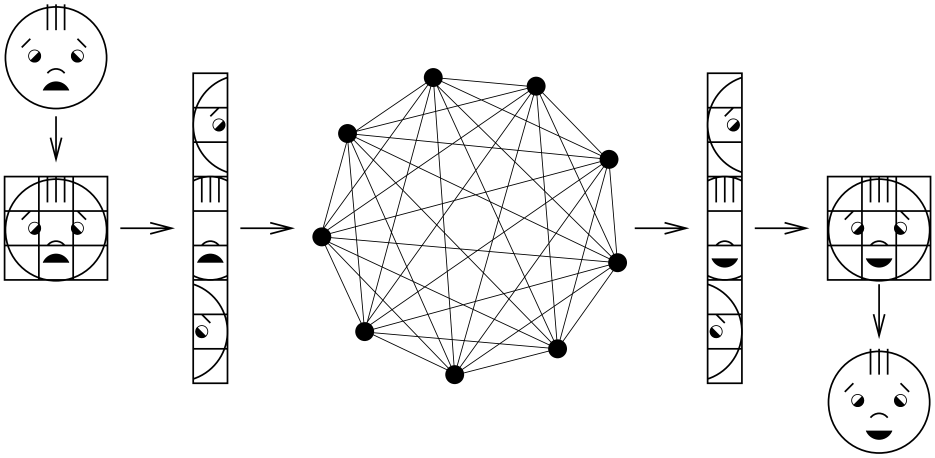

FIGURE 2. The testing phase of an autoassociative memory. The memory units are initialized to the gray-level values of a digitized memory cue. The activation level of each unit is calculated by summing the weighted values of all connected units. The resultant unit values are reformed into a face image.

0.) In this case, which applies to face images, reconstruction performance can be improved by using Widrow-Hoff learning to create the connection weights.

The Widrow-Hoff learning rule creates connection weights through iterative error-correction cycles; each cycle composed of learning and testing. The test portion of a cycle is accomplished by the reconstruction process previously described. First, the difference between the test response and the desired output is calculated. Connection weights are then modified by adding an adjustment proportional to the product of the amount of error and the activation level of their units. This procedure can be shown to be equivalent to using a Least Mean Square algorithm (see [2] for details.) The amount of correction applied is controlled by a small positive learning constant, which, when chosen correctly, makes the amount of error decrease over repeated cycles, until the response matches perfectly the target. This iterative procedure is performed for each of the patterns to be learned. Although Widrow-Hoff learning is more suited than Hebbian learning to the demands of this particular model, such may not always be the case (see [14] for a detailed comparison of these two techniques in the face processing domain.)

2.1.3. *Autoassociative Memory - Principal Component Analysis Approach*

An autoassociative memory can also be implemented by decomposing the statistical structure (i.e., eigen-decomposition) of the pattern cross-product matrix [4, 9]. This is commonly called the principal component analysis (PCA) approach, first applied to face images in a neural network framework by Abdi [1], and used in these simulations. In PCA, the weight matrix is decomposed as a sum of elementary constituents called “principal components” or eigenvectors (sometimes eigenfaces in the context of face models.) Each eigenvector of the pixel cross-product matrix can be considered as a kind of “feature” of, or way of describing, the faces. Unlike traditional face features (e.g., eyes, nose, and mouth), which are localized within the face, the eigenvector “features” span the entire face [13]. To each eigenvector is associated an eigenvalue, which gives the amount of pattern variance attributable to each eigenvector. The eigenvectors are typically ordered according to their eigenvalue: the one explaining the most variance being called the first eigenvector, the one explaining the second most variance being called the second eigenvector, and so on. If the eigenvalues are all made equal to unity, the equivalent of full or “perfect” Widrow-Hoff learning is achieved.

A pattern, such as a face image, can be reconstructed from a multiplication of the eigenvectors and eigenvalues of the cross-product matrix times the face vector. It has also been shown that a face can be approximately reconstructed by using only a subset of the eigenvectors, those with the largest eigenvalues [13]. Moreover, different ranges of eigenvectors have been found best suited for different face tasks: eigenvectors with large eigenvalues being optimal for general categorization tasks (in particular the second eigenvector for determining the sex of a face) and eigenvectors having small eigenvalues being optimal for face recognition [10].

2.1.4. *Perceptron Classifier*

The perceptron, first proposed by Rosenblatt [12] to model perceptual activities, is a simple classifier. In neural network terms, its most basic version consists of one input layer of units (sometimes called the retina) with weighted connections to one output layer. To assign an input stimulus to one of only two categories, the output layer consists of a single unit. The activation of this unit, calculated as the sum of all weighted inputs, is transformed into a binary response (e.g., If the activation is greater than 0 the response is 1, otherwise the response is 0.) Each response represents one of the two categories. Connection weights are established so as to create linear decision boundaries, which separate the input vectors into categories.

An alternative, and equivalent, perspective considers the perceptron as determining the distance of a given stimulus from a prototype for each of the categories. A category prototype can be represented as a vector of the same dimension as the input layer vector by calculating the center of gravity (barycenter) of a set of representative stimuli. The Euclidean distance of an input vector to each of the prototype vectors is calculated. The closest prototype is taken as the classification of the stimulus.

In our model, the projections of both learned and novel faces on the eigenvectors of the learned faces provide the input for the sex classification decision. The face projection onto an eigenvector is a scalar indicating how much that eigenvector contributes to that face structure. Male and female prototypes are created by averaging the projections of all learned male and female faces respectively. Then, the projections of a test face onto the eigenvectors of the learned faces are calculated.

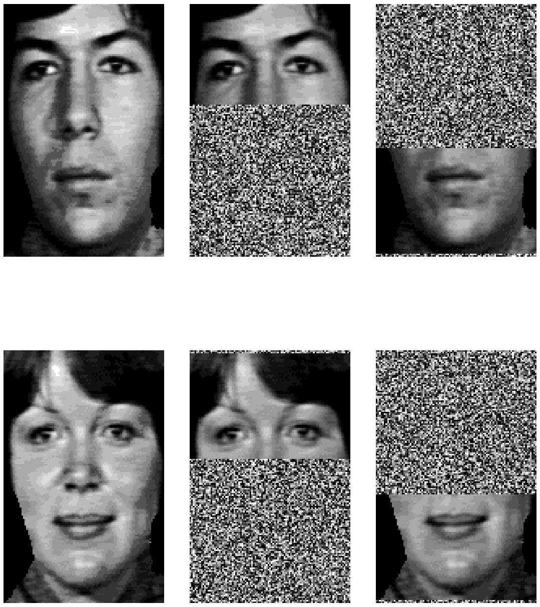

FIGURE 3. The three stimulus conditions for a male face (top row) and a female face (bottom row): 1) full face with hair masked, 156 × 99 = 15444 pixels (left), 2) top of face, 65 × 99 = 6435 full image pixels retained (center), 3) bottom of face, 65 × 99 = 6435 full image pixels retained (right).

Finally, the Euclidean distance of these projections from each of the two prototypes is computed, with the shortest distance determining the test face classification.

2.2. ***Method***

2.2.1. *Stimuli*

The simulation stimuli were pictures of 160 faces (80 male and 80 female), digitized using 16 gray levels. These pictures contained no distinguishing characteristics, such as jewelry, glasses, or facial hair, and were aligned roughly at eye level. Each face was prepared under three conditions, as illustrated in Figure 3: 1) full face image cropped at the top and sides, and further masked to cover hair, 2) the top portion only of Condition 1 stimuli, including forehead, eyes, and the upper part of the nose, 3) the bottom portion only of Condition 1 stimuli, including the jaw area from below the nose to the chin. An equal number of facial area pixels were retained for Conditions 2 and 3, and the remainder of the image was masked with random values. Thus, the retained facial area for the top and the bottom conditions remained positionally aligned with the full face image.

Stimulus	All Faces		Male Faces		Female Faces	
Condition	Learned	Novel	Learned	Novel	Learned	Novel
1 - Full Face	1.000	.775	1.000	.800	1.000	.750
2 - Top	.860	.694	.920	.763	.800	.625
3 - Bottom	.825	.663	.838	.688	.813	.638

TABLE 1. Proportion of correct model classifications by stimulus condition and learning condition, over all faces, male faces only, and female faces only.

2.2.2. *Procedure*

The simulations were all based on training the model with full face images, and testing the classification ability with novel images from all three stimulus conditions.

To test the ability of the model to classify both learned and novel stimuli, a bootstrap technique was used as described below. The memory was trained with Condition 1 stimuli for 60 males and 60 females, and then tested on the images that had been learned, and also on the other 20 males and 20 females that had not been learned. This procedure was done four times, with a totally different group of 40 unlearned faces used for testing on each iteration. This set of four iterations was done once for each stimulus condition.

2.3. ***Results and Discussion***

Table 1 shows the proportion of correct sex classification obtained by the model under the three stimulus conditions, for both learned and novel faces. Because the stimulus condition for learning was Condition 1 (full face with hair cropped) classification of the learned faces in this condition was always perfect.

Several points can be noted from these results.

- The model always performs above chance level.
- Although the model always performs better for learned faces, it shows some ability to generalize learning by an overall classification accuracy for novel faces ranging from 66% to 78%.
- Male faces are better classified than female faces under all conditions. Because neither the learning nor testing stimuli included the hair, this superior classification of male faces cannot be attributed to the variance of the female hair styles, negating the prior conjecture based on full face stimuli with hair included. In fact, calculation of the variance for the face stimulus in Conditions 1, 2, and 3 shows greater variance for the female images under all these conditions.
- Overall classification accuracy of novel faces declines from stimulus Condition 1 (77.5%) to Condition 2 (69.4%), and also from Condition 2 to Condition 3

Stimulus Condition	1 Full Face	2 Top	3 Bottom
1 Full Face		$r_{\mathrm{m}} = 0.45529$ $p = 0.0001$	$r_{\mathrm{m}} = 0.40452$ $p = 0.0002$
2 Top	$r_{\mathrm{f}} = 0.26833$ $p = 0.0161$		$r_{\mathrm{m}} = 0.25745$ $p = 0.0211$
3 Bottom	$r_{\mathrm{f}} = 0.40534$ $p = 0.0002$	$r_{\mathrm{f}} = -0.15442$ $p = 0.1714$	

TABLE 2. Model classification correlations between the three stimulus conditions for individual male and female faces. The upper right triangle shows correlations for the male faces, and the lower left triangle shows correlations for female faces.

(66.25%). The difference in accuracy between the top and bottom condition is relatively small.

- The pattern of difference in accuracy for Conditions 1, 2 and 3 is not the same for male (80%, 76.3%, 68.75%) and female (75%, 62.5%, 63.75%) faces. Although male faces yielded a close to linear decline from Condition 1 to 2 to 3, the accuracy for female faces dropped more substantially from Condition 1 to Condition 2, and hardly differed at all between Conditions 2 and 3.

To determine if there was a relationship between the model classification for the same face under different stimulus conditions, we calculated the correlations shown in Table 2. These correlations between classification for the same novel faces, across the three simulated stimulus conditions, are highest for male images between the full face and top of face conditions, and lowest for female images between the top of face and bottom of face conditions.

In summary, the specific predictions of the model, to be tested against human subject performance, are as follows.

- Male faces will be better classified than female faces under all conditions.
- The major decline in overall performance will be between the full face and the partial face conditions, and the classification accuracy in the top and the bottom conditions will be roughly equivalent.
- The pattern of classification accuracy will not be the same for male and female faces. Male faces will yield a close to linear decline from full face to top to bottom conditions, but the accuracy for female faces will drop more substantially from the full face to the top condition, and be roughly equivalent between the top and the bottom condition.
- The correlations between classification for the same novel faces will follow the pattern exhibited in Table 2 and be highest for male images between full face and top conditions, and lowest for female images between top and bottom conditions.

3. Experiment

The purpose of this experiment was to gather human data on accuracy in classifying the sex of whole or partial face images, for both male and female faces. These data were then compared to the predictions of the model. Because of our interest in determining if the model has the predictive power to sex classify the same stimuli, in a manner analogous to human subjects, the experiment was designed to support an item analysis of human performance on particular faces, under varying conditions of presentation.

Considering the high level of human expertise typically exhibited in this type of classification (e.g., 96% accuracy with no make-up and hair covered [5]), very high performance was expected from human subjects on full face classification, even with the hair masked. This condition serves mainly to ensure that the face stimuli used here are generally comparable to those used in other studies, and to provide a baseline for comparison with the model results. Dividing the face image into top and bottom portions roughly divides the features considered to be major contributors to sex classification (eyes and jaw), and also disrupts the total configuration. However, unlike presenting features in isolation, some configural information is maintained for each part of the face. So, the stimuli for this experiment do not closely match any of the stimuli used in the human subject experiments mentioned in the introduction of this paper, but are somewhat of a combination. If the facial areas of eye and jaw are similarly salient for sex classification when viewed out of total context, as when viewed in the total face context, human subject accuracy under the top and bottom conditions should not be strikingly different. This is what the model has predicted. However, if human ability for this task is strongly based on the relation of certain areas to the total configuration of the face image, disruption may be greater under one of the conditions.

3.1. *Method*

3.1.1. *Subjects*

Sixty University of Texas at Dallas undergraduate students (17 male and 43 female) participated in the experiment in exchange for a core psychology course research credit.

3.1.2. *Stimuli*

The experimental stimuli were the same faces, and parts of faces, used in the simulations (as illustrated in Figure 3), with the exception that the images were displayed on a black background. The display did not include the random values that were added to provide the mask for the simulations.

3.1.3. *Experimental Design*

The experimental results were analyzed as a three factor mixed design: $\mathcal{S}_{20}(\mathcal{A}_3) \times \mathcal{B}_{80}(\mathcal{C}_2)$. The between subjects factor $\mathcal{A}$ was the condition of the face stimuli with three levels. The within subjects factors were the face presented ($\mathcal{B}$, a random factor with 80 levels) nested within the sex of the face ($\mathcal{C}$, male or female). The dependent variable was the accuracy of the sex classification.

3.1.4. *Procedure*

Subjects were randomly assigned to one of the three stimulus conditions. The assignments were blocked, with each subject in a group of three consecutively tested subjects assigned randomly to one of the three conditions. Upon arrival, subjects were given written directions for the experiment, explaining that they would see a face, or portion of a face, briefly displayed on the monitor, and then be asked to decide "male" or "female" by clicking the mouse on the appropriate screen area. There was no time pressure for this decision. The face stimuli were presented, and the subject responses recorded, by a MATLAB program on a Sun SPARCstation LX running Unix Solaris. Images of the same 160 faces, formatted in the assigned condition, were briefly displayed against a black background in a random order to each subject. After one second had elapsed, the image was covered by a decision mask. When the subject had responded, the next face image was displayed.

3.2. ***Results and Discussion***

A mixed ANOVA[2] showed all main effects and interactions to be significant at the $\alpha = .01$ level. This is not surprising considering the large numbers of degrees of freedom in this design. These effects are presented and discussed below, grouped by major areas of interest.

3.2.1. *Main Effects of Sex of Face and Stimulus Condition*

The main effect of sex of face, $F'(1,\ 172) = 6.88$, $MS_e = 1.6$, $p < .01$, shows that, across the three conditions, subjects performed better when classifying male images (88% mean accuracy) than when classifying female images (82% mean accuracy). The proportion of variance explained by this effect, however, is very small (.009). We shall return to discuss this proportion, in comparison with other proportions of variance, later on. The left panel of Figure 4 compares the subject accuracy by sex of image to the predictions of the model. Although model accuracy was lower for both sexes, the shape of the results was well predicted. Both human subjects and the model classified male faces better than female faces by about the same amount.

A second main effect of particular interest is that of stimulus condition, $F'(2,\ 253) = 61.83$, $MS_e = .65$, $p < .01$. As shown in the right panel of Figure 4, mean sex classification accuracy declined as the amount of facial area presented at test decreased. However, although the proportion of the face image retained for Conditions 2 and 3 was the same, the task difficulty increases more substantially when the bottom portion of the face is presented, as contrasted to the top portion.

[2] Because the faces were treated as a random factor in the design, a quasi-F was calculated for all effects that did not include the face factor.

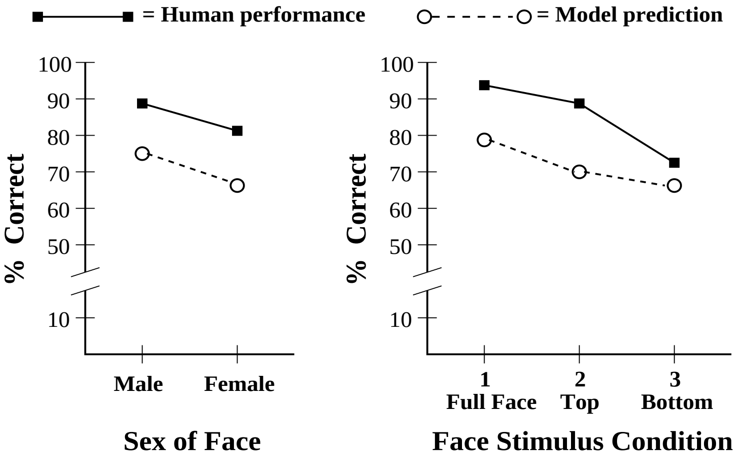

FIGURE 4. Main effects of sex of face and stimulus condition compared to model predictions. The left panel shows the mean performance accuracy for human subjects by sex of face (male faces $M = .88$, $SE = .005$, $N = 4800$; female faces $M = .82$, $SE = .006$, $N = 4800$) above the respective model predictions (male faces = .75, female faces = .67). The right panel shows the mean performance accuracy across stimulus conditions for the human subjects: 1) $M = .94$, $SE = .004$, $N = 3200$; 2) $M = .89$, $SE = .006$, $N = 3200$; 3) $M = .72$, $SE = .008$, $N = 3200$; above the respective model predictions: 1) .78, 2) .69, 3) .66.

The right panel of Figure 4 compares the subject accuracy for the three stimulus conditions to the predictions of the model. Although the subject and model accuracy both decline across these conditions, it can be seen that the model predictions are not as good for this effect as they were for the main effect of sex of face. The model did not predict the sharper decline in accuracy between the top and the bottom condition shown by human subjects.

3.2.2. *Interaction between Stimulus Condition and Sex of Face*

Unexpectedly, the difference in accuracy between stimulus conditions is not the same for male and female face images. Figure 5 shows this interaction between the stimulus condition and the sex of the face, $F'(2,\ 134) = 26$, $MS_e = 1.15$, $p < .01$. The interaction is predominantly characterized by a strong decline in accuracy from the top of the face to the bottom of face for female images (94% and 59% respectively) as compared to a modest increase in accuracy for male faces (83% and 86% respectively). However, it should be noted that the proportion of total variance explained by this interaction is less than 5% ($R^2 = .049$).

Figure 6 compares the model predictions for Conditions 1, 2, and 3 to the human subject results. The model predictions for Conditions 2 and 3 are not supported

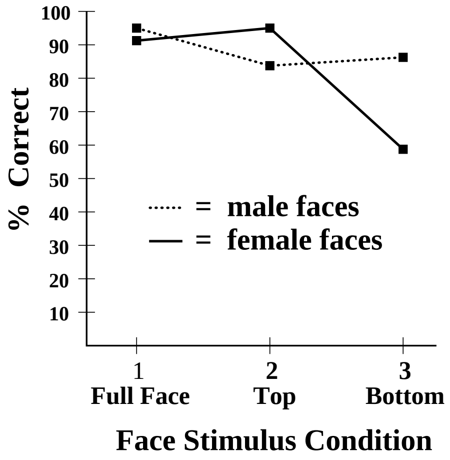

FIGURE 5. Mean human subject performance accuracy by sex of face image across three stimulus conditions:
1) for male faces $M = .96$, $SE = .005$, $N = 1600$;
for female faces $M = .92$, $SE = .007$, $N = 1600$;
2) for male faces $M = .83$, $SE = .009$, $N = 1600$;
for female faces $M = .94$, $SE = .006$, $N = 1600$;
3) for male faces $M = .86$, $SE = .009$, $N = 1600$;
for female faces $M = .59$, $SE = .012$, $N = 1600$.

by the human data, particularly for female faces. Note that, in this one case, the model outperforms the human subjects.

Why was the classification of the bottom of a face more difficult for human subjects on female face stimuli than on male face stimuli? There are several possible explanations for this difficulty. One, of course, is that the picture quality was simply poorer for the lower portion of the female faces. However, this seems unlikely when the performance of the model is considered. It may be that the jaw and mouth, seen in this partial context of half of a face, are not as good cues for female faces as for male faces. Most subjects assigned to Condition 3 mentioned upon completion that they thought the task difficult. See how well you do classifying the stimuli shown in Figure 7. An intriguing conjecture is that the absence of any hair surrounding the jaw led subjects to adopt a male bias when they were uncertain of the sex. The model, having no "hair experience" whatsoever, could not be biased in this manner. Finally, we questioned if the sex of the subjects (predominantly female) could be contributing to this interaction.

3.2.3. *Effect of Sex of Subjects*

As noted previously, most of the human subjects in this experiment were female (43 out of 60). The experimental design did not include the sex of the subjects as a factor, and no attempt was made to balance the number of male and female

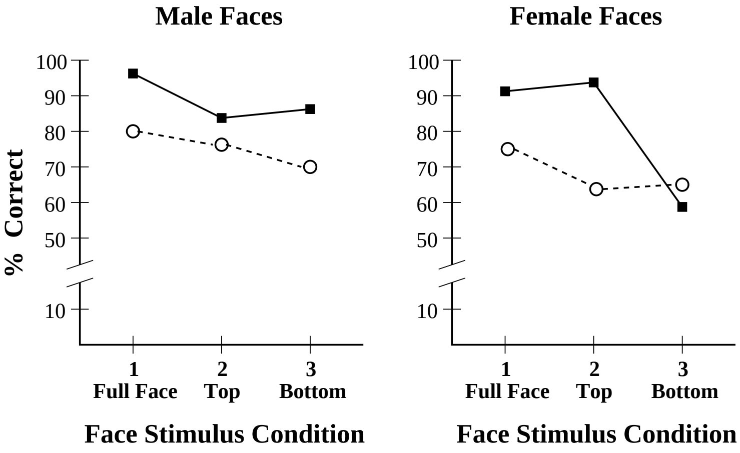

FIGURE 6. Comparison of human subject sex classification accuracy and model predictions by sex of face. The left panel compares data for male faces by stimulus condition:
1) subjects $M = .96$, model = .80;
2) subjects $M = .83$, model = .76;
3) subjects $M = .86$, model = .69.
The right panel compares data for female faces by stimulus condition:
1) subjects $M = .92$, model = .75;
2) subjects $M = .94$, model = .63;
3) subjects $M = .59$, model = .64.

subjects assigned to each twenty subject condition (there were four male subjects in Condition 1, six in Condition 2, and seven in Condition 3). However, after the fact, we considered the possibility that the surprisingly poor accuracy found for female faces in Condition 3 could be related to the predominance of female subjects. To determine if this was the case, we compared the correct response frequency of male subjects for male and female faces to the correct response frequencies of female subjects. Chi-square tests did not show any difference in accuracy between male and female subjects and sex of face over all conditions, $\chi^2 = .005$, $N = 8153$, $p = .944$, or for stimulus condition × sex of face: for Condition 1, $\chi^2 = .133$, $N = 2996$, $p = .715$; for Condition 2, $\chi^2 = 1.986$, $N = 2844$, $p = .159$; for Condition 3, $\chi^2 = .278$, $N = 2313$, $p = .598$. Thus, even with large total frequencies, no relationship between the sex of subjects and accurate classification of male and female faces was found.

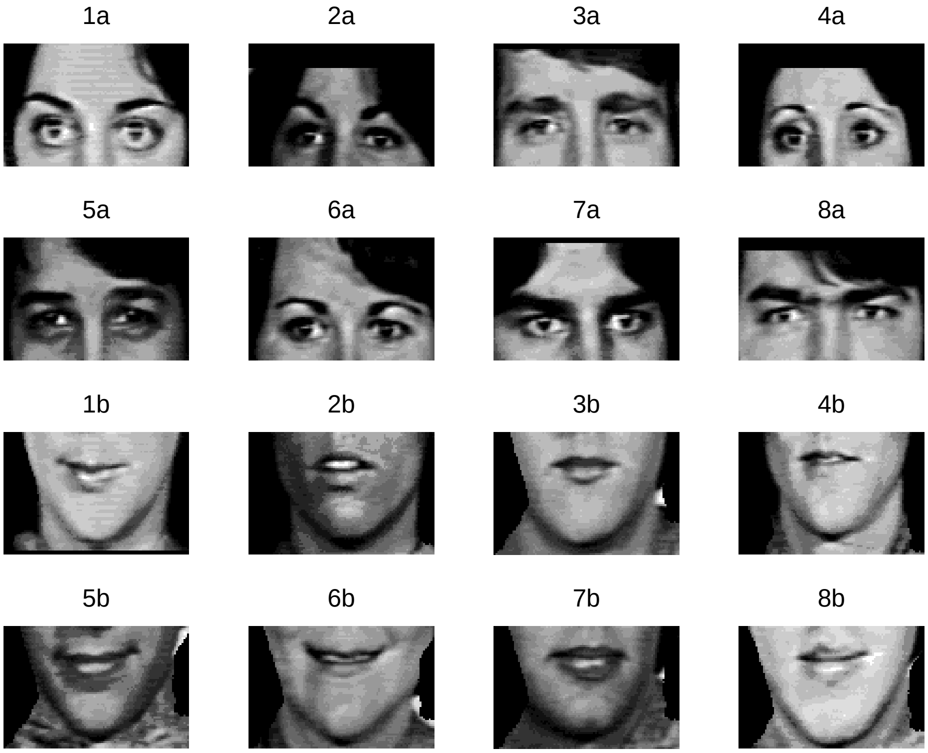

FIGURE 7. Sample classification test stimuli. Classify each image and see how you do.
Answer: Stimuli with numbers 1, 2, 4, and 6 are female; 3, 5, 7, and 8 are male. The stimuli having the same number are parts of the same face (i.e., 2a and 2b are the same person.)

3.2.4. *Item Effect of the Faces*

The data also revealed a main effect of face nested in sex of face, $F(158,\ 9006) = 12.22$, $MS_e = .08$, $p < .0001$, showing a strong item effect. Classification performance varies depending upon the particular faces judged within one sex. Particular faces were also differentially classified depending upon the condition of their presentation as shown by an interaction effect between stimulus condition and face, $F(316,\ 9006) = 6.51$, $MS_e = .08$, $p < .0001$.

A correlation analysis of the scores given to the same faces under the three stimulus conditions (see Table 3) provides some additional information about the importance of different facial areas for sex classification. Note the difference between the correlations for male and female faces. Correlations between the various stimulus conditions for male faces vary from $r = .26$ to $r = .49$. In contrast, the highest correlation, $r = .62$, is found between Conditions 1 (full face) and 2 (top of face) for

Stimulus Condition	1 Full Face	2 Top	3 Bottom
1 Full Face		$r_{\mathrm{m}} = 0.4943$ $p = 0.0001$	$r_{\mathrm{m}} = 0.2978$ $p = 0.0073$
2 Top	$r_{\mathrm{f}} = 0.6157$ $p = 0.0001$		$r_{\mathrm{m}} = 0.2627$ $p = 0.0185$
3 Bottom	$r_{\mathrm{f}} = 0.2364$ $p = 0.0347$	$r_{\mathrm{f}} = -0.0749$ $p = 0.5094$	

TABLE 3. Human subject classification correlations between three stimulus conditions for male and female faces. The upper right triangle shows correlations for the male faces, and the lower left triangle shows correlations for female faces.

female faces, and the lowest correlation, $r = -.07$, is also found for female faces, and occurs between the top of the face and the bottom of the face. Therefore, although the classification given to a female full face with hair cropped is a fair predictor of the classification given to the top of the same face, the top of face classification shows no linear relationship to the classification given to the bottom of the face. A comparison of these correlations with those obtained from the simulations (see Table 2) shows a fair similarity of pattern for the male faces. However, for female faces, the model performance is definitely contrary to that of the human subjects, with the exception of predicting no correlation between the classification of the top and bottom stimulus conditions.

3.2.5. *Proportion of Variance Explained*

Figure 8 shows the proportion of variance explained by each effect in the design. The total variance explained by the experimental effects is only 42.3%. Although no one effect accounts for a majority of this variance, it is informative to examine the contributions of these proportions. Summed together the effects that include the face variable (Face, Condition × Face) do account for a majority of the explained variance. This proportion lends support to the proposition that item effects are considerable when using faces as stimuli, pointing out the necessity of replicating any experiment of this type with different face stimuli (or treating the face stimuli as a random factor.) It is possible that very different results would be obtained using different faces. Lesser amounts of variance are explained by the stimulus condition (7%) and the interaction of the condition and the sex of the face image (5%). The proportion of variance that can be ascribed to the sex of the face alone (less than 1%) is, in fact, very small.

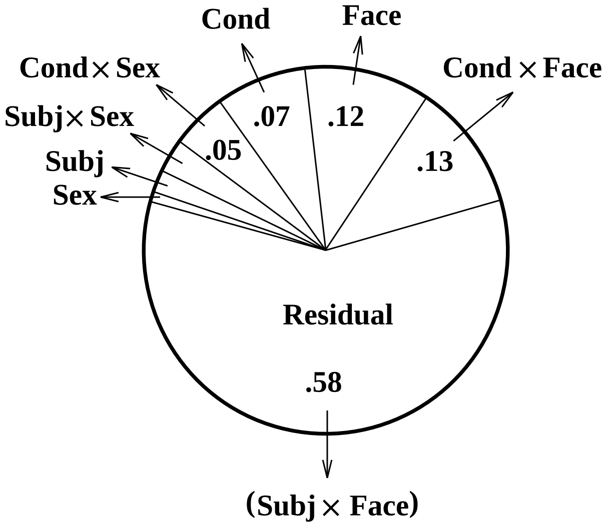

FIGURE 8. Proportion of variance explained by the experimental effects. The label sex indicates the sex of the face image.

3.2.6. *Item Analysis*

The model may demonstrate some predictive power if it can indicate which faces are most likely to be classified correctly by human subjects, and which faces are most likely to be missed. To determine if this is the case, we compared the average number of human subjects correctly classifying the stimuli correctly classified by the model, to the average number of human subjects correctly classifying the stimuli misclassified by the model. Figure 9 illustrates this comparison. It can easily be seen that, for male faces, the model has no such predictive power. The best tendency for prediction occurs only for female faces in the full face and top of face conditions.

A final analysis compared the classification decisions made for particular face images by the model to those made by the human subjects. Even though the model was 78% correct on the full face condition (Condition 1), compared to 94% correct for the human subjects, the classifications assigned to particular faces by the model are not at all well correlated with the classifications given by individual subjects. For this condition (which yielded the best model performance), this correlation ranges from a high of $r = .19$, $p < .02$ to a low of $r = -.09$, $p = .278$, with $r = .12$ on average. The amount of similarity for individual face classification between the model and human subjects declines even more dramatically for the other conditions. For Condition 2, the correlations with human subjects ranges from a high of $r = .13$, $p = .10$ to a low of $r = -.09$, $p = .28$, with $r = .019$ on average. For Condition 3, this correlation ranges from $r = .10$, $p = .21$ to $r = -.12$, $p = .14$, with an average of $r = .003$.

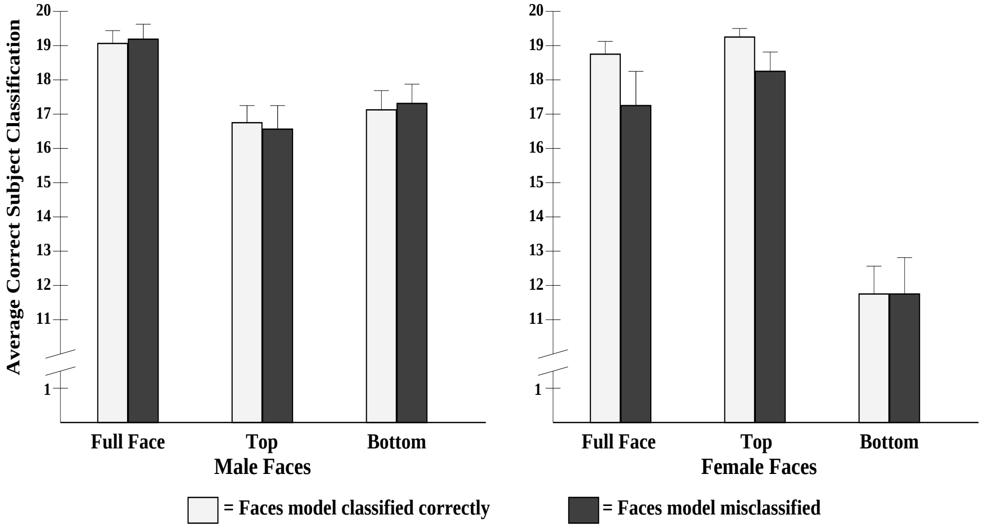

FIGURE 9. Average number of subjects out of 20 who correctly classified male and female faces in the three stimulus conditions for: 1) faces correctly classified by the model (gray bars) and 2) faces misclassified by the model (black bars). The number of face images used to calculate the average subject score varies by condition, sex of face, and model classification. The error bars represent the standard error of the means.

Figure 10 illustrates the item difference for faces between the model predictions and human performance by comparing the predictive power of the model for each face to the predictive power of a particular subject for the same face under each stimulus condition. The selected subject is the one who scored closest to the mean score for that condition. Correlations of the classification given to each of the 160 faces by this "average" subject, and by the model, with the classifications given by the other 19 subjects for these same faces in the same experimental condition, are shown in the figure in descending order of the simulation correlation. The vertical lines in the figure, which connect the model and "average" subject correlations with the same other subject, show the discrepancy between these two correlations.

For Condition 1, Figure 10 shows that the model correlation with the subjects never exceeded .20 and the average correlation was $r = .12$, $p = .1377$. Also, the correlation pattern predicted does not match the pattern of the "average subject" very well. For the most part, the model is not classifying the same faces in the same way as the human subjects. Note however, the "average" subject is not a very good predictor either. The correlation of classification by face of the selected subject in Condition 1, with the other subjects in the same condition, never exceeds .53 and has an average of $r = .34$, $r^2 = .11$. The "average" subject correlations with other

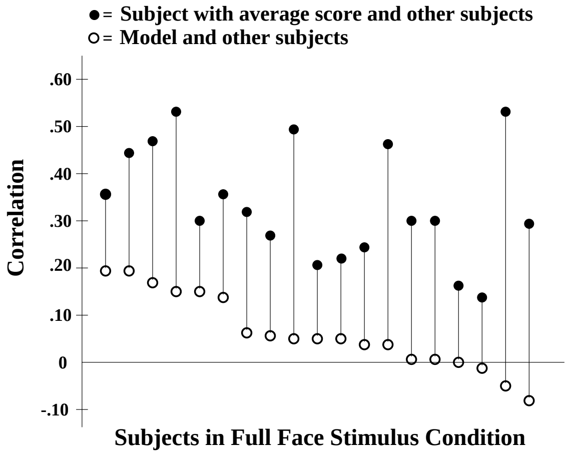

FIGURE 10. Correlations of sex classifications by face between a human subject with average accuracy and each of the other 19 subjects in the same stimulus condition (filled circles), and the correlations of the classifications predicted by the model and the same 19 human subjects for the same faces (hollow circles). The correlations of the "average" subject and the model, with the same human subject, are connected with vertical lines, and arranged from left to right in descending order of model correlation.

subjects in Condition 2 do not exceed .40, and on average $r = .25$, $r^2 = .06$. The correlations for Condition 3 do not exceed .35, and on average $r = .24$, $r^2 = .06$.

4. Is There a Quick Fix?

Is there any way to improve the predictive power of this model, without sacrificing simplicity? It has been suggested to us that the lack of predictive power on an item level might be attributable to the classification approach. By using the perceptron, we are comparing novel stimuli to the averages of the learned male and female faces. This simulates the formation of a gender prototype by human subjects. If we consider classification by human subjects to be based on the similarity of the novel stimulus to learned examples, the psychological relevance of the model might be improved by using a nearest neighbor classifier. Therefore, we conducted some additional simulations using a very simplistic exemplar approach. In these preliminary simulations, the projection of a novel face was compared to the projections of all the learned faces, and classified according to the sex of the closest matching face.

Unfortunately, this alternate approach did not produce any predictive advantage. In fact, overall accuracy for novel stimuli declined substantially. For example, in the full face condition classification accuracy declined from 78% to 66%. This approach outperformed the perceptron in only one respect: It was amazingly accurate (99%) when tested with the top portion only of the *learned* faces. Thus, it appears that, at least in this model, the perceptron is substantially superior for classifying novel stimuli: The nearest neighbor approach does not yield the same ability to generalize learning. Of course, these preliminary results do not rule out the possibility of predictive improvement through the use of other measures of similarity, but only indicate that a more complex mechanism may be necessary.

Other avenues for further investigation lay in the representation of the faces. It is possible that the use of projections on all eigenvectors captured too much information, including details specific to individual faces, which do not relate to this general categorization task. Therefore, simulations that use only eigenvectors with the highest eigenvalues, particularly the second eigenvector (see [10]), might produce better results. Differential weighting of facial representation, such as face areas and/or eigenvectors should also be considered.

5. Conclusion

In summary, the human subject data did not support the predictions of the model. Although, the main effect of male images classified better than female images was predicted, this effect was not constant over all stimulus conditions for human subjects. An interaction between sex of face and stimulus condition showed better human subject performance for female faces than for male faces in Condition 2 (top of the face), but substantially poorer performance for female faces in Condition 3 (bottom of the face). In contrast, the model had predicted a close to linear decline across the three stimulus conditions for male faces, and for female faces a substantial drop between Conditions 1 and 2, and relative equivalency for Conditions 2 and 3. Also, the patterns of the correlations of classifications given to the same faces across conditions were not the same for the model and the human subjects. However, the model did predict the lack of correlation between the top and the bottom of the face for female images.

The item analysis indicates that the model is not classifying the faces in the same way as human subjects. Correct model classification did not predict at all the likelihood of correct classification by subjects for male faces. For female faces there was some tendency shown for this predictive power in Conditions 1 and 2, but this did not extend to Condition 3. However, the classifications given to a particular face are not well correlated between the human subjects either, showing that the human subjects are not performing in the same manner as each other.

What could account for the differences between the sex classifications made by the model and those of the human subjects, and also between the subjects themselves? One obvious explanation might be that the technique used in the model has very little, or perhaps no, commonality with the abilities and decision criteria of most human subjects. However, recall that the subjects themselves did not perform the task in a uniform manner either. The difference in face experience is another possible cause for the model's failure to predict. Although the approach taken in the simulations attempted to give the model memory “experience” somewhat analogous to that of human beings, the experience given was, of course, not the same. Neither the amount, nor the sample, of face training provided to the model approaches the human experience. Moreover, the model never “experienced” hair. This might have contributed to the model classification accuracy surpassing that of the human subjects for the female faces in Condition 3. Without any “knowledge of hair,” the absence of hair could not influence the model toward a male classification, as might have occurred for the human subjects.

Perhaps a closer look at the human data can explain some of the problems of the model. The proportion of variance explained by the significant experimental effects showed the strongest effects were the item effect of the faces and the interaction of the faces with the stimulus conditions. Whereas, the greatest contribution to the variance of results (almost 60%) was the interaction between the face stimuli and the subjects. In other words, not only do the stimuli show a strong item effect, and the subjects exhibit differences, but these two variables interact producing major subject differences across the faces. Therefore, it seems unlikely that any one model, at least one that does not incorporate some intrinsic variability or maybe random components, could predict the human behavior.

Finally, the performance of the human subjects in Conditions 2 and 3, is of particular interest, and should be further investigated using different face stimuli. The results obtained from this experiment, using these particular faces, indicate that, for female face images only, the eye region, including eyes, brows, and forehead, is a better cue for sex classification than the mouth, chin, and jaw, when seen in a partial face context. Whereas, for male faces, the lower portion of the face is a slightly better cue than the eye region. The model, which simply used an analytical comparison of the statistical structure of the faces, did not show the large difference for female faces between the top and the bottom condition, as shown by human subjects. Therefore, factors other than a simple application of statistical structure must contribute to the human behavior. Whether these factors are a consequence of the facial structure itself, the experience of the subjects, other biases, or some combination of all of these and more, remains open for further investigation.

References

[1] Abdi H, A generalized approach for connectionist auto-associative memories: Interpretation, implication and illustration for face processing. In *Artificial intelligence and cognitive sciences*, ed. by Demougeot J, Herve T, Rialle V and Roche C (Manchester University Press, Manchester, 1988) pp. 151–164.

[2] Abdi H, A neural network primer. *Journal of Biological Systems* **2** (1994) 247–281.

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

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

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

[6] Brown E and Perrett DI, What gives a face its gender? *Perception* **22** (1993) 829–840.

[7] Brunelli R and Poggio T, HyperBF networks for sex classification. In *Proceedings of the Image Understanding Workshop, DARPA, San Diego* (Morgan Kaufman, San Mateo, CA, January 1992) pp. 311–314.

[8] Fleming MK and Cottrell GW, Categorization of faces using unsupervised feature extraction. In *Proceedings of the International Joint Conference on Neural Networks, Vol II* (San Diego, CA, June 1990) pp. 65–70.

[9] Kohonen T, *Associative memory: A system theoretical approach* (Springer Verlag, Berlin, 1977).

[10] O'Toole AJ, Abdi H, Deffenbacher KA and Valentin D, Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America A* **10** (1993) 405–411.

[11] Roberts T and Bruce V, Feature saliency in judging the sex and familiarity of faces. *Perception* **17** (1988) 475–481.

[12] Rosenblatt F, The perceptron: a probabilistic model for information storage and organisation in the brain. *Psychological Review* **65** (1958) 386–408.

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

[14] Valentin D, Abdi H, Edelman BG and Nijdam A, Connectionist "face"-off: Different algorithms for different tasks. *Psychologica Belgica* **36** (1996) 65–92.

[15] Valentin D, Abdi H, Edelman B and O'Toole AJ, Principal component and neural networks analyses of face images: What can be generalized in gender classification? *Journal of Mathematical Psychology* in press.

[16] Yamaguchi MK, Hirukawa T and Kanazawa S, Judgment of gender through facial parts. *Perception* **24** (1995) 563–575.

Appendix A. Simulation Model Detail

Notations

Let $\mathbf{x}_k$ be the I-dimensional vector representing the k^{th} face out of a set of K faces. The complete set of faces can be described by an $I \times K$ matrix denoted $\mathbf{X}$.

Hebbian Learning

Hebbian learning is achieved by taking the sum of the outer products of each input with itself, in matrix notation:

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

where T indicates the transpose function, which reverses row and column. The formula above has the effect of creating $\mathbf{W}$ with elements $w_{i,j}$ such that each element is the sum from all input patterns of the product of unit i and unit j. Usually the input vectors are first normalized for computational convenience, so that their inner product is equal to 1 (i.e., $\mathbf{x}_k^T \mathbf{x}_k = 1$).

Widrow-Hoff Learning

The Widrow-Hoff algorithm, also called the delta rule, develops the connection matrix by first initializing $\mathbf{W}$ to all zeros and then iteratively adding increments denoted by $\Delta\mathbf{W}$:

$$\mathbf{W}_{[n+1]} = \mathbf{W}_{[n]} + \Delta_n \mathbf{W}$$

where n represents an iteration step. The increment $\Delta_n \mathbf{W}$ is a function of the difference between an input and its reconstructed output:

$$\Delta_n \mathbf{W} = \eta(\mathbf{x} - \mathbf{x}_{[n]})\mathbf{x}^T$$

where η is a small positive learning constant, $\mathbf{x}$ represents the original stimulus, and $\mathbf{x}_{[n]}$ represents the answer state of the system at time n. As time steps increase, the output will gradually become more similar to the input, and so $\Delta\mathbf{W}$ will decrease, tending toward zero. The algorithm will ultimately reach convergence if η is chosen properly. The iteration described is accomplished in turn for each pattern to be "learned."

Reconstruction

If the connection weights are arranged in a matrix having a row and a column for each unit, and a memory key is expressed as a column vector, retrieval is accomplished by multiplying the weight matrix by the pattern vector.

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

Principal Component Analysis Approach

The actual calculation approach used for the model, and detailed below, is one described by Valentin, Abdi, Edelman, and O'Toole [15]. Due to the large size of the matrix of faces to be learned ($I \times K$), where $I = 15,444$ is the number of pixels per face and $K = 120$ is the number of faces, it is not computationally feasible to obtain the eigenvectors of the cross-product matrix directly. For this reason, these eigenvectors and eigenvalues were obtained by singular value decomposition (SVD). This decomposition transforms the matrix of the faces to be learned, $\mathbf{X}$, into two sets of eigenvectors and one set of corresponding eigenvalues. One set of eigenvectors, denoted $\mathbf{U}$, is from the pixel cross-product matrix ($\mathbf{X}\mathbf{X}^T$); the other set of eigenvectors, denoted $\mathbf{V}$, is from the face cross-product matrix ($\mathbf{X}^T\mathbf{X}$). The eigenvalues pertain to both sets of eigenvectors. This decomposition is such that the face matrix $\mathbf{X}$ may be reconstructed by:

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

where $\boldsymbol{\Delta}$ is the diagonal matrix of the square roots of the eigenvalues.

In general, the projections, denoted $\mathbf{P}$, of face vectors in $\mathbf{X}$ onto $\mathbf{U}$ can be calculated as:

$$\mathbf{P}_{\text{face}} = \mathbf{X}^T\mathbf{U} \ .$$

However, to accomplish complete Widrow-Hoff learning, which sphericizes the weight matrix (all non-zero eigenvalues are equal to one), the projections of the learned faces were calculated as:

$$\mathbf{P}_{\text{learn}} = \mathbf{X}^T_{\text{learn}}\mathbf{U}\boldsymbol{\Delta}^{-1} = \mathbf{V}$$

and the projections of the test faces as

$$\mathbf{P}_{\text{test}} = \mathbf{X}^T_{\text{test}}\mathbf{U}\boldsymbol{\Delta}^{-1} \ .$$

Perceptron

The classification perceptron was implemented by first finding the average of the projections on the eigenvectors for the learned male faces and for the learned female faces:

$$\bar{\mathbf{p}}_{\text{male}} = \frac{1}{N}\sum^{N} \mathbf{p}_{\text{learned male}}$$

$$\bar{\mathbf{p}}_{\text{female}} = \frac{1}{N}\sum^{N} \mathbf{p}_{\text{learned female}}$$

where N is the number of faces learned for each sex. The Euclidean distance of the projections of each test face from both the average of the learned male face projections and the learned female face projections was calculated as:

$$d(\mathbf{p}_{\text{test}}, \bar{\mathbf{p}}_{\text{male}}) = \|\mathbf{p}_{\text{test}} - \bar{\mathbf{p}}_{\text{male}}\|$$

$$d(\mathbf{p}_{\text{test}}, \bar{\mathbf{p}}_{\text{female}}) = \|\mathbf{p}_{\text{test}} - \bar{\mathbf{p}}_{\text{female}}\| \ .$$

The sex of the averaged projections that gave the smallest distance was chosen as the sex classification of the test face.