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

MANIPULATING FACE GENDER: EFFECTS ON CATEGORIZATION AND RECOGNITION JUDGMENTS

KENNETH A. DEFFENBACHER*, CHERYL HENDRICKSON*, ALICE J. O'TOOLE†, DAVID P. HUFF†, AND HERVÉ ABDI†

* *Department of Psychology, University of Nebraska at Omaha, Omaha, NE 68182-0274, USA*
and
† *School of Human Development, GR4.1, The University of Texas at Dallas, Richardson, TX 70083-0688, USA*

ABSTRACT

Previous research has shown that faces coded as pixel-based images may be constructed from an appropriately weighted combination of statistical "features" (eigenvectors) which are useful for discriminating members of a learned set of images. We have shown previously that two of the most heavily weighted features are important in predicting face gender. Using a simple computational model, we adjusted weightings of these features in more masculine and more feminine directions for both male and female adult Caucasian faces. In Experiment 1, cross-gender face image alterations (e.g., feminizing male faces) reduced both gender classification speed and accuracy for young adult Caucasian observers, whereas same-gender alterations (e.g., masculinizing male faces) had no effect as compared to unaltered controls. Effects on femininity-masculinity ratings mirrored those obtained on gender classification speed and accuracy. We controlled statistically for possible effects of image distortion incurred by our gender manipulations. In Experiment 2 we replicated the same pattern of accuracy data. Combined, these data indicate the psychological relevance of the features derived from the computational model. Despite having different effects on the ease of gender classification, neither sort of gender alteration negatively impacted face recognition (Experiment 3), yielding evidence for a model of face recognition wherein gender and familiarity processing proceed in parallel.
Keywords: Face gender, face recognition, face images, facial features, neural networks.

1. Introduction

Though there is a clear evolutionary advantage to being able to classify quickly a face as male or female and to recognize it as belonging to a familiar person, it is not clear whether these two processes are independent. Ellis [12, 13] has proposed a hierarchical model of face recognition in which familiarity processing is a third stage following a face classification stage (Is it a face?) and a "visually-derived semantic" categorization stage (gender, age, race) [8]. However, Bruce, Ellis, Gibling, and Young [7] obtained results challenging Ellis' theory: Faces whose gender was more difficult to classify were no harder to recognize as familiar. Bruce et al. [7] argued for a model wherein gender and familiarity processing proceed in parallel, with gender judgments typically completed first. They did point out, however, that to the extent that gender judgments shared information sources with familiarity judgments, that one could expect to find evidence for a perceptual hierarchy. Bruce et al. advocated further research efforts to identify sources of shared and unshared information and to specify computationally "the nature of the representations derived from these different sources."

A number of recent investigations have been concerned with specifying the information sources for gender judgments [5, 6, 9, 23, 27]. Both Brown and Perrett [5] and Roberts and Bruce [23] have concluded that, although observers are able to extract gender information from isolated, explicit facial features and their additive combination (e.g., eyes and brows), these sorts of cues probably function ordinarily in a configural manner, in relation to one another and to global face shape. To this list of explicit features, Bruce et al. [6] have added skin texture and overall three-dimensional structure as gender-relevant sources of information. Finally, Burton et al. [9] have implemented a computational model of gender categorization. They first computed an elaborate set of facial measures, simple distances between key points of facial features (e.g., interocular distance), ratios and angles formed between these points (e.g., mouth width to mouth-nose distance), and three-dimensional distances derived from full-face and profile views. Burton et al. then used these measures as input to a discriminant function analysis. Results were mixed, performance of the model with raw distance measurements on full-face photographs (85% accuracy) being well below that of human observers (96% accuracy); performance of the model with combined full-face photograph and three-dimensional measurements (94% accuracy) approached that of human observers, though 16 variables were required. Burton et al. commented on the sheer difficulty of finding a set of explicit facial feature measures that permit human levels of gender categorization performance.

A rather different approach to specifying computationally the nature of the representation of face gender information is that taken by various statistically-based models (often implemented as connectionist or neural network models). A number of these models have been shown to be capable of classifying sets of faces by gender and/or race [see 25, for a thorough review]. While these models can operate on virtually any sort of face representation, including the kinds of feature-based measures used by Burton et al. [9], they have most commonly been used in conjunction with relatively unprocessed data, raw, pixel-based data taken from face images [e.g., 18] separated two-dimensional shape and image data [14], and separated three-dimensional shape and image data from laser scans of human heads [21]. The primary strength of these kinds of models applied to relatively unprocessed data is that they bypass the problems associated with the a priori selection of the task relevant features. Whether implemented via a connectionist algorithm or statistical analysis, these models are directed at extracting, via example, the "features" that are most useful for the task to which they are applied. The question which then arises is whether the information or features used by these models for gender classification is psychologically salient.

INSERT FIG 1 AROUND HERE

Our current work was motivated by the results of a number of studies wherein we and others have presented normalized, pixel-based face representations to a linear autoassociative network [cf. 14, 16-18]. The network performs a principal components analysis (PCA) on a cross-product matrix of pixel values made from the entire set of face images and hence is a formal approach to analyzing the statistical structure of the face set. Using a principal components based representation, individual faces can be represented as weighted combinations of eigenvectors extracted from the cross-product matrix of learned face images. These eigenvectors can be displayed as images and generally appear face-like; they can be thought of as *macrofeatures* [2], inasmuch as they are the statistical features from which a face is constructed (see Fig. 1 for an example). Depending on the set of faces learned by the network, some eigenvectors have been shown to relate to visually-derived semantic dimensions such as race [15] or gender [15, 16]. For instance, O'Toole et al. [16] found the weight on the second eigenvector ("amount" of this principal component required to reconstruct a particular face) to be the best predictor of face gender. By adding and subtracting this eigenvector from the first eigenvector, it could be seen that it captured information relevant for contrasting male and female hair styles and also for contrasting global differences in the shapes of male and female faces, with female faces having rounder and fleshier faces than male faces. O'Toole et al. [16, 18] also found that eigenvectors with relatively small eigenvalues, those components explaining small amounts of variance, provided better information for face recognition. Intuitively this makes good sense in that the information most useful for recognizing a face is the information that the face least shares with other faces – i.e., what makes the face unique or different from other faces. This information is likely to be captured by eigenvectors with relatively small eigenvalues, since they explain only small proportions of variance in the face set.

The PCA enables us to quantify the information in faces that is most useful for making a gender judgment. The first question we consider in the present experiments concerns the psychological relevance of this information. Specifically, in Experiments 1 and 2 we wish to determine if these model-predicted sources of gender information relate to gender classification speed and accuracy of human observers, as well as to the perceived masculinity/femininity of faces. We address this question by manipulating synthetically the model-predicted sources of gender information in individual faces, and by reassessing human observers' gender classification performance on the faces. We show that manipulating the model-predicted gender information does indeed affect the gender classification performance of observers in expected ways, in at least some experimental conditions. The success of these manipulations allowed us to make use of the synthetically altered faces for addressing more theoretically based issues concerning the relationship between familiarity and visually-based classification processes. In Experiment 3, we carry out a standard recognition memory study with the gender-altered faces to determine whether manipulations of the information useful for gender classification impact judgments of whether a face has been seen before. The latter question bears on the issue of whether gender processing is hierarchically related to familiarity processing.

INSERT FIG 2 AROUND HERE

In the first two experiments, we present three sorts of male and female faces to observers, *normal faces* (NORM), those reconstructed with no alterations to eigenvector weights, *"masculinized"* faces (MAS), reconstructed with second and third eigenvector weights altered in a masculine direction, and *"feminized" faces* (FEM), reconstructed with second and third eigenvector weights altered in a feminine direction (see Fig. 2 for examples). At this point we need to show that these two eigenvectors are indeed important in coding gender for the faces used in our experiments. Figure 3 displays a scattergram wherein we have plotted the standardized (z-score) projections on the second and third eigenvectors for each of the male and female faces used in these experiments. Interestingly, 75% of the female faces are above the antidiagonal (i.e., the diagonal going from bottom left to top right), and 79% of the male faces are below it. Female faces tend to have negative values on the second eigenvector (74% of cases) and positive values on the third (61% of cases). Male faces, on the other hand, tend to have positive values on the second eigenvector (86% of cases) and negative values on the third (56% of cases). Thus male and female faces tend to have contrasting values on each of the eigenvectors, with the second eigenvector discriminating more sharply than the third, as O'Toole et al. [16] found.

INSERT FIG 3 AROUND HERE

Now given the presence of hair length differences amongst our faces and given that the second eigenvector contains information relevant to gender differences in hair style, it might be conjectured that gender coding of our faces is a simple matter of tracking differences on the dimension of hair length. At one level, this issue is simply a matter of determining the extent to which our observers might rely on hair as a cue to face gender in the present experiments. We will address this matter shortly. At a second, more philosophical level, the relevant question is whether or not one should consider "hair" cues as a valid part of the information in human faces. This is a rather different question that has generated a good deal of debate among face researchers for some time now. It is perhaps worth beginning with the relevant facts upon which most researchers are likely to agree. First, under normal, ecologically valid circumstances, most people have hair, and would feel rather awkward being seen in public without it. Second, when hair is available as a recognition or gender classification cue, observers will indeed make use of it. In fact, at least for recognition of unfamiliar faces, observers find external features such as hair shape more useful than internal features for discriminating among faces on a recognition memory task [11]. Finally, it is impossible to eliminate hair from a face without eliminating a good bit of head surface, which many researchers would consider a very valid part of the information in faces.

For a gender classification task, however, there are two related issues that must be considered, ones having to do with the fact that hair is changeable via styling and that hair styles are dictated by cultural norms that often differ for men and women. One assumes, of course, that the internal features of a face are completely free of such cultural cues. This is perhaps not a safe assumption, however. Currently, for example, eyebrow plucking and shaping is a fashionable trend, one popular with women, but not with men. Hence, it is at least clear that even the internal features of faces also may contain cues to gender that might be compared to hair style. Should eyebrows therefore be eliminated from facial stimuli? It would be neither surprising nor unadaptive if people became atuned to useful cultural cues to face gender and made use of them in face processing tasks. For this reason, we consider the cultural cues to be a valid part of the information in faces, which observers make use of to the degree that they are informative.

Next we address the more specific question concerning the role of hair cues in the present experiments. Is gender coding of our faces simply a matter of tracking differences in hair length, as defined by the second eigenvector? We would make two points in reply. First, the second eigenvector also contains information relevant to gender differences in global face shape (see Fig. 2, again). Second, as the scattergram in Figure 3 makes evident, the third eigenvector likewise contributes to the coding of gender differences [cf. 16], and this eigenvector is unrelated to hair length. O'Toole, Deffenbacher, Valentin, McKee, Huff, and Abdi [19] have shown that weight of a face on the third eigenvector is related to the reaction time to classify it as male or female. Thus there is evidence that hair length, as captured in the second eigenvector, is unlikely to provide a complete account of effects seen in the present experiments.

We predicted that cross-gender face alterations should produce decreased gender classification speed and accuracy in a speeded classification task and lowered ratings of masculinity and femininity as compared to faces in the other two conditions, unaltered faces and those with same-gender alterations ("super males and super females"). Conversely, same-gender alterations should produce faces that are classified with increased speed and accuracy and that are rated as more masculine and feminine than unaltered faces or those with cross-gender alterations.

2. **Experiment 1**

In the first experiment, four measures were taken from observer data. Reaction times and errors were taken from the first task, a speeded gender classification task. A second task required observers to rate the femininity-masculinity of faces. Due to the fact that the synthetic image alterations, in some cases, produced faces that appeared distorted (see Fig. 2, again), we wished to measure the degree of face degradation caused by our gender manipulations, so that we could partial out this source of variance. Therefore, we asked observers to rate the face images for naturalness of appearance, providing thereby, a perceptual measure of face distortion, which we could use to partial out the changes in performance that might result with the gender-altered faces.

2.1. ***Method***

2.1.1. *Participants and Design*

The 24 Caucasian observers (14 female and 10 male) were volunteers from the University of Texas at Dallas community and ranged from 20 to 52 years of age. Student volunteers enrolled in undergraduate psychology courses received one research credit for participating. The initial experimental design was a three-factor mixed analysis of variance (ANOVA), with observer gender as a between-subjects factor and type of face gender manipulation (NORM, FEM, or MAS) and face gender (male or female) being within- subjects factors.

2.1.2. *Stimuli*

The X Windows tool described by O'Toole and Thompson [20] was used to generate three sets of 144 face images (72 male and 72 female). These faces were generated as a sum of 160 eigenvectors from an autoassociative memory originally constructed from 160 Caucasian faces (80 male and 80 female). The three face sets included a set of normal faces, those reconstructed with no alteration to eigenvector weights, a set of masculinized faces, reconstructed with their second and third eigenvectors each weighted at $+50$, and a set of feminized faces, reconstructed with their second and third eigenvectors each weighted at -50. These weight changes correspond to the units used in the public domain eigenvector tool described in O'Toole and Thompson [20]. The X Windows tool values are simply 100 times the true values, and thus a value of 50 corresponds to a weight of .50. We wish to be clear that we make no metric claims about this manipulation. These tool values were chosen by trial and error so as to produce the greatest amount of exaggeration consistent with the least amount of distortion possible. We should also note that in the context of eigenvector analysis, the sign of the weights is completely arbitrary (i.e., no distortion of the face space would occur by rotating the axes). Each face was a 225×151-pixel grayscale image with 16 gray levels per pixel (see Fig. 2, again).

As noted previously, PCA has been applied, in very recent work, to analysing a variety of different face codings. It is perhaps worth discussing here, briefly, how the present coding relates to these others. In addition to the analysis of raw image-based codes, PCA has been applied to face images that have been separated into a two-part representation consisting of: a.) the two-dimensional configuration of their facial landmarks (i.e., corners of the eyes, etc.); and b.) their shape-free image intensity information [10, 14]. The former shape-based coding is simply a set of interfacial landmark distances, whereas the latter shape-free code is made by morphing the face images to a standard configuration. This alignment process represents important progress in the quality of the stimulus that is available to the PCA for analysis, in that it eliminates problems with face parts that do not always coincide in the pixel-based representations that are aligned only informally, (usually by eye height and forehead center). However, due to the limited number of facial landmarks that are typically used in these analyses, 34, for instance [14], complex and subtle information about face shape (e.g., cheek contour) is not retained in the shape- based coding, though it may be captured somewhat in the image code. PCA has also been used to analyze the separated three-dimensional structure and graylevel image data available from laser scans of human heads [21]. While this lacks the advantages associated with facial landmark correspondence, it retains very complex shape information about the entire structure of the head. A good compromise to these two approaches is being explored in recent work aimed at solving a pixel-based correspondence problem for face images [22, 26]. While good progress has been made on this problem, these algorithms are computationally very intensive and are not yet sufficiently well-developed or specified in the literature to allow for comprehensive testing on a reasonably natural set of stimuli.

In the present study, we have worked on face images aligned only by eye height and by the center point between the eyes. While this approach is computationally less sophisticated than those mentioned above, there is good evidence from psychological data that PCA applied to a raw, image-based representation captures information that relates to human recognition memory performance on faces and to human observers' facial ratings [18].

2.1.3. *Procedure*

Necessary counterbalancing was carried out such that although each observer was presented with 24 male and 24 female faces in each of the three gender manipulation conditions, a particular face was seen by each observer at only one level of manipulation. Further, across observers, each face was seen by an equal number of observers at each level of manipulation. Subject to these restrictions, a different random order of 144 face images was created for each observer. The first task was a speeded gender classification task wherein each face image was presented for 250 ms, followed by a 100-ms, 50% gray mask. Observers responded "male" or "female" as quickly as possible by pressing one of two labeled keys on a computer keyboard. After a 1-min break, observers were presented with the 144 face images, again, and in self-paced fashion, rated each image first for femininity-masculinity on a 1 (very feminine) to 10 (very masculine) scale, and then for naturalness of appearance on a 1 (very natural-looking image) to 10 (very distorted-looking image) scale. Both sets of ratings were accomplished by pressing number keys on a computer keyboard.

2.2. *Results*

In order to control for variance due to some of the altered face images appearing distorted, we performed linear regression analyses to determine how well naturalness ratings predicted observers' reaction times, errors, and femininity-masculinity ratings. The coefficients of determination from the regression analyses were $r^2(142) = .03$, $p < .05$, for reaction time, $r^2(142) = .22$, $p < .001$, for classification errors, and $r^2(142) = .00$ for femininity-masculinity ratings. Residuals were then computed on these measures (differences between actual and predicted values) and used as the dependent variables in subsequent ANOVAs. Observer gender showed no significant main effects or interactions in any of the ANOVAs. Thus, we collapsed our experimental design across observer gender, yielding a 3 (levels of gender manipulation) $\times$ 2 (face gender) within subjects design.

2.2.1. *Reaction Time and Accuracy*

Figure 4 shows observers' unregressed reaction times to classify the face images by gender. Reaction times are plotted for ease of interpretation, because the pattern of mean residuals was virtually the same. Analysis showed only a significant interaction between level of gender manipulation and face gender, $F(2, 46) = 12.97$, $MS_e = 10132.22$, $p < .001$. A planned comparison indicated that cross-gender face alterations (FEM male faces and MAS female faces) produced significantly slower reaction times as compared with the faces in the other two gender manipulation conditions, $F(1, 46) = 10.32$, $MS_e = 21879.46$, $p < .005$.

INSERT FIG 4 and 5 AROUND HERE

Error rates for the gender categorization task are displayed in Figure 5, unregressed error rates in Figure 5(a) and residual error rates in 5(b), given that the latter showed a somewhat different pattern of means. Analysis revealed both a significant interaction between gender manipulation and face gender, $F(2, 46) = 71.16$, $MS_e = 0.0156$, $p < .001$, and a significant main effect of gender manipulation, $F(2, 46) = 6.20$, $MS_e = 0.0124$, $p < .005$. A planned comparison showed that cross-gender face alterations produced significantly higher error rates as compared with the faces in the same-gender alteration and no alteration conditions, $F(1, 46) = 73.24$, $MS_e = 0.0198$, $p < .001$. A post hoc comparison showed that there was not a significant difference in error rates when categorizing normal female versus normal male faces.

INSERT FIG 6 AROUND HERE

2.2.2. *Femininity-Masculinity Ratings*

Figure 6 displays observers' unregressed ratings of femininity-masculinity for the face images. An analysis of variance applied to these rating scale data [3] demonstrated that male faces at every level of gender manipulation were rated as more masculine than female faces, $F(1,23) = 721.30$, $MS_e = 0.8271$, $p < .001$. In addition, there was a significant interaction of gender manipulation and face gender, $F(2,46) = 98.43$, $MS_e = 0.2062$, $p < .001$, as well as a main effect of gender manipulation, $F(2,46) = 92.79$, $MS_e = 0.2917$, $p < .001$. A planned comparison showed that cross-gender face alterations generated appropriately different femininity-masculinity ratings, with male faces in the FEM condition rated more feminine than male faces in the other two conditions, and female faces in the MAS condition rated more masculine than the other sorts of female faces, $F(1,46) = 15.65$, $MS_e = 2.2559$, $p < .001$. Unexpectedly, however, the male faces in the MAS condition were rated as more feminine than normal, unaltered male faces, $F(1,46) = 12.46$, $MS_e = 2.2551$, $p < .001$ (post hoc comparison).

2.3. *Discussion*

The pattern of results obtained in this experiment allow us to answer our first empirical question in the affirmative. Indeed, certain model-predicted sources of gender information were related to gender classification speed and accuracy of human observers in predictable ways. Specifically, cross-gender alterations of second and third eigenvector weights produced decided slowing of classification speed, on the order of 110-120 ms. These same alterations produced a decided decrease in accuracy of gender classification, too, as much as a 42% increase in absolute error rate. Further, both sorts of gender alterations together generated a generally predictable pattern of femininity-masculinity ratings, a pattern mirroring that in the reaction time and error rate data.

The finding that faces with greater cross-gender weighting on the second or third eigenvector slowed the speed of gender classification is not without precedent. In the study of O'Toole, Deffenbacher, Valentin, McKee, Huff, and Abdi [19] female faces with more masculine values of the second eigenvector weight tended to be classified more slowly as female.

Somewhat surprising was the fact that increases in same-gender weightings on these two eigenvectors did not have effects opposite to those of increases in cross-gender weightings. Increased same-gender weightings did not create "super-male" or "super-female" faces. Except for one anomalous finding (cf. Fig. 6, male faces in the MAS condition), the faces with same-gender alterations were perceived as if they were within the normal range of variation in cue values for their gender.

These results appear, at first glance, to be at variance with a result of Benson and Perrett [4]. Using a facial morphing algorithm, Benson and Perrett blended 16 female student faces to create an "average" female student face; this process was repeated with 16 male faces. They then created a "hyper-male" and a "hyper-female" face by doubling the differences between each feature point in the "average" male and female faces. Thus, for example, if the "average" male nose were longer than that of the "average" female, the difference in length would have been doubled in the process of creating the "hyper-male" face. In creating the "hyper-female" face, the same difference in nose length would have been doubled so that the new nose length would have been twice as many units shorter than that of the "average" male nose. Benson and Perrett reported that the "hyper-male" face was rated as more masculine than the prototype male face. They did not report whether the "hyper-female" face was rated as more feminine than the prototypical female face they had created. For a similar procedure applied to male and female Japanese faces, see [27]. While in the present study, hyper-male and hyper-female effects were not found, there are a number of important differences in the kind of algorithm used by Benson and Perrett and ourselves. Primarily, Benson and Perrett's algorithm distorts the mean of a set of faces, whereas the present procedure exaggerated the gender relevant information in individual faces. Additionally, their morphing algorithm operates on the two-dimensional configuration of the features of the face. Our approach, by contrast, captures image-based information in the individual faces that relates to face gender.

One other finding is interesting, if unexpected. As Figure 6 makes clear, unaltered (NORM) male faces were perceived as more masculine (3.8 scale points from the midpoint of the femininity-masculinity scale) than unaltered female faces were perceived as feminine (1.7 scale points from the scale midpoint). Perhaps a partial explanation of this result is discussed by Abdi, Valentin, Edelman, and O'Toole [1], in a study wherein they used the same faces as in the present experiment. We would cast the argument this way: The male faces in Experiment 1 may well have been more similar to one another than were the female faces. That is, they may have been more tightly clustered around the center of male face space than the female faces were around the center of female face space. Abdi et al. [1] indeed found that some female faces were actually more similar to male faces than to other female faces.

3. **Experiment 2**

The primary purpose of this experiment was to replicate the basic pattern of results in Experiment 1, while also adopting a different approach to dealing with any problems of apparent "unnaturalness" in the face images. We reasoned that if we shortened exposure duration of each image, then any distortion or grotesqueness would be less apparent. We accomplished this by presenting the face images for 50 ms each, followed by a 200-ms mask. At this brief duration, our focus was on accuracy of gender categorization, and thus a speeded response was not emphasized. We also incorporated a direct measure of image quality in this study, which we used to assess the results, parceling out image distortion as a potential cause of the gender manipulation effect. Finally, we wanted to ensure that Experiment 1's pattern of results was not due in any way to the fact that type of face image was not blocked for stimulus presentation. In the present experiment, unaltered and altered face images were blocked.

3.1. ***Method***

3.1.1. *Participants and Design*

Participants were 22 Caucasian undergraduate volunteers (13 females and 9 males) from the same university community as that of Experiment 1's participants. They were exposed to all stimulus presentation conditions of a 3 (level of gender manipulation) $\times$ 2 (face gender) within-subjects ANOVA design.

3.1.2. *Stimuli*

The unaltered images were 150 faces (75 of each gender) generated as a sum of eigenvectors from the same autoassociative memory of 160 Caucasian faces used in Experiment 1. We should note that the original set of faces were all of young adults, without facial hair or glasses, photographed in front of a homogeneous light background. After digitizing, the images had been aligned so that the eyes of all faces were at about the same height and so that the center point between the eyes was at the same place in all photographs. The faces had not been normalized explicitly for size. However, since the photographs had been taken at the same camera distance, faces were roughly equal in size.

Gender alteration was accomplished in the same manner as Experiment 1. Two presentation blocks of images were created. One block included all 150 unaltered face images. The other included 150 altered face images, randomly selected from a pool of 300 gender-altered faces (75 FEM females, 75 FEM males, 75 MAS females, and 75 MAS males).

3.1.3. *Procedure*

All participants were tested individually. Each observer was presented with the two blocks of 150 face images for 50 ms per image, followed in each instance by a 200-ms mask, an evenly textured dot pattern. The order of presentation blocks was counterbalanced across observers. Observers categorized each face as male or female by pressing one of two computer keys. As mentioned previously, a speeded response was not emphasized.

INSERT FIG 7 AROUND HERE

3.2. ***Results and Discussion***

Mean gender classification error rates as a function of the six experimental conditions are plotted in Figure 7. As in Experiment 1, an ANOVA of these data revealed a main effect for gender manipulation, $F(2, 42) = 123.60$, $MS_e = 99.78$, $p < .001$, and an interaction of gender manipulation with face gender, $F(2, 42) = 98.77$, $MS_e = 347.24$, $p < .001$. Unlike Experiment 1, however, there was also a main effect of face gender, $F(1, 21) = 7.51$, $MS_e = 414.51, p < .05$. We conducted simple effects analyses of the interaction effect in order to test for face gender differences at each level of gender manipulation. For unaltered (NORM) faces, there was no significant difference between the mean error rate for female faces (11.36%) and the mean for male faces (7.77%). In the case of feminized (FEM) faces, however, there was indeed a significant effect ($p < .01$) of face gender, with the error rate for male faces being much higher (77.55%) than that for female faces (6.55%). Finally, for masculinized faces (MAS), there was a statistically reliable effect ($p < .01$) of face gender, with the error rate for female faces (52.0%) being greater than that for male faces (13.73%).

Additionally, as in Experiment 1, we wished to apply a statistical control for the role of face naturalness in these effects and also wished to include a measure of the image quality with reference to the unaltered images. Perhaps the simplest measure of image quality is the cosine between the unaltered (original) and altered faces. This is akin to the correlation coefficient that measures the similarity of the unaltered and altered faces.[1] We were also curious to see if this measure related to the observer naturalness ratings assessed in Experiment 1. We correlated the six gender manipulation by face gender means for the naturalness ratings with the cosine means for these conditions and found excellent agreement, $r(4) = -.96$, $p < .01$. Thus, observer ratings of naturalness seemed to be capturing objective aspects of the image degradation. That is, gender-altered faces with high cosines (low image degradation) were regularly associated with low ratings on our scale (natural-looking faces).

Next, we found that the observer accuracy data for the gender manipulation and face gender groups was indeed reliably predicted by the cosine measure for these conditions, $r^2(130) = .43$, $p < .001$. To assess the importance of this correlation in accounting for our effects, we repeated the ANOVA, using as the dependent variable the residuals computed from the linear regression predicting accuracy from cosine. This yielded results qualitatively identical to those reported above, with a highly reliable main effect of gender manipulation, $F(2, 42) = 37.77$, $MS_e = 99.78$ $p < .001$, a main effect of face gender $F(1, 21) = 8.00$, $MS_e = 414.51$, $p < .05$, and an interaction of gender manipulation with face gender, $F(2, 42) = 41.90$, $MS_e = 347.24$, $p < .001$. Thus, the effects of image degradation were generally not sufficient to account for observers' differential performance as a function of face gender and gender manipulation.

Thus in Experiment 2 we were able to replicate the same pattern of results for the error rate data as that obtained in Experiment 1. Same-gender face-image alterations did not reduce gender classification accuracy, but cross-gender alterations (FEM males and MAS females) did. We should note that this replication was obtained despite using a different means of attempting to minimize any effects on our results of face image distortion. Hence, the procedure of making distortion less obvious to

[1] The cosine measure was the cosine of the angle between two vectors of pixel values, one for the unaltered face and one for the altered one. In each case, these vectors consisted of the concatenation of the rows of the matrix of pixel values comprising the facial image. Strictly speaking, the cosine of two vectors is their inner product divided by the product of their lengths. If both vectors have zero mean, the cosine reduces to the familiar coefficient of correlation.

the observer through shortened presentation time was apparently as effective as the statistical control used in Experiment 1.

The only difference between the results of Experiments 1 and 2 concerns absolute levels of gender classification error obtained. Interestingly, even despite the considerably lower stimulus duration in Experiment 2, the absolute error rate for gender classification remained virtually the same as in Experiment 1 for the unaltered face images and those with same-gender alterations. Indeed, two-tailed t tests confirm the lack of significant experimentwise differences in error rate for these four stimulus conditions. Only the absolute error rates for the two cross-gender conditions were adversely affected by the briefer presentation time of Experiment 2, drastically so, increasing by a factor of approximately two-thirds over comparable rates in Experiment 1. We should like to suggest that for faces whose gender cues fall within the normal range of variation for their gender, a relatively accurate decision as to their gender is based on cues that can be extracted by the visual system from a presentation as brief as 50 ms.

4. **Experiment 3**

In this experiment, we attempted to answer our second empirical question. Does manipulation of face gender cues impact recognition memory for these same faces? We have found that cross-gender cue manipulations negatively impact the speed (Experiment 1) and accuracy (Experiments 1 and 2) of gender categorization. If the cross-gender manipulations were to produce faces no more difficult to recognize as familiar than face images without such manipulation, then we should have found support for Bruce et al.'s [7] arguments against the notion of a perceptual hierarchy in face processing. By contrast, if only cross-gender manipulations were to produce faces more difficult to recognize as familiar, we should have found support for Ellis' [12, 13] hierarchical model of face processing, wherein familiarity processing (third stage) can only proceed after visually-derived semantic categorization along the gender dimension has been completed (second stage).

4.1. ***Method***

4.1.1. *Participants and Design*

A total of 48 (24 male, 24 female) Caucasian undergraduate students from the University of Nebraska at Omaha participated in this experiment. All students received course extra credit in exchange for their participation. One group of 24 participants (12 male, 12 female) was assigned to a 2 (observer gender) $\times$ 2 (face gender) $\times$ 2 (gender manipulation: NORM vs. FEM) mixed ANOVA design. The other group of 24 observers was assigned to another $2 \times 2 \times 2$ mixed ANOVA design that differed only in the particular levels of gender manipulation, NORM vs. MAS. Observer gender was explicitly included in the design, again, because of the somewhat inconsistent findings in the face recognition memory literature regarding its effects [24].

4.1.2. *Stimuli*

Sixty target faces and 60 distractors were selected from the same 160-image set as before. Recognition memory tasks were presented to two separate groups of 24 undergraduates. One group's targets comprised 15 male and 15 female NORM faces and 30 FEM faces, 15 of each gender; their distractor set was similarly constituted. The other group saw 30 NORM and 30 MAS faces in both target and distractor sets.

4.1.3. *Procedure*

All observers were tested individually. After reading instructions presented by microcomputer, observers first viewed each of 60 target faces for 5 s each and a 3-s interstimulus interval (ISI). Immediately following this study phase a computerized message alerted observers to a 1-min break. Then, after reading test instructions, observers were presented with 120 faces (60 targets, 60 distractors) for 5 s each with a 5-s ISI. Observers were asked to indicate as quickly and accurately as possible, whether each test face was "old" (target face) or "new" (distractor face) by pressing one of two computer keys. Each observer received a different random order of study and test faces. In addition, for half the observers, target and distractor face sets were reversed. Thus each face image was presented both as a target and as a distractor.

4.2. **Results**

4.2.1. *Feminization of Faces*

For each of the 24 observers in the group exposed to feminized faces, four d' scores were computed, one for each of the four stimulus conditions, NORM male, NORM female, FEM male, FEM female. Each d' score was calculated from the hit rate obtained on 15 target faces and the false alarm rate obtained on the 15 distractor faces of the identical type. Thus the score for an observer on FEM female faces was based on his/her hit rate on 15 target FEM female faces and false alarm rate on 15 distractor FEM female faces. As an additional control for the effects of image degradation, we used as the dependent variable in the ANOVA, the residual scores from a linear regression, in this experiment, predicting observer d' from the image quality (i.e., cosine) measure associated with a particular experimental condition. In this instance, the coefficient of determination was $r^2(94) = .04$, $p < .05$.

Submitting these scores to the $2 \times 2 \times 2$ ANOVA mentioned previously, yielded no significant effects. It is worth noting that an ANOVA on the raw d' scores (i.e., not the residual scores) did yield a main effect of gender manipulation, $F(1,22) = 5.32$, $MS_e = .69$, $p < .05$, but no interaction of gender manipulation and face gender. Recognition memory in this instance was less accurate for FEM faces ($d' = .68$) than for NORM faces ($d' = 1.07$). In combination, these two ANOVAs indicate that the gender manipulation per se did not affect recognition accuracy.

4.2.2. *Masculinization of Faces*

Four d' scores were again computed for each of the 24 observers in the separate groups exposed to masculinized faces, scores for NORM male, NORM female, MAS male, and MAS female. Scores were based on the same number of target and distractor faces as before. The $2 \times 2 \times 2$ ANOVA likewise was the same as before, save for the obvious change in the gender manipulation factor to include MAS vs. NORM face images. Additionally, as the dependent variable in the ANOVA, we again used the residual of a linear regression predicting d' from the cosine measure, $r^2(94) = .02$, $p > .05$.

Submitting these scores to the $2 \times 2 \times 2$ ANOVA mentioned previously, again, yielded no significant effects. In this case, also, an ANOVA on the raw d' scores yielded only a main effect of gender manipulation, $F(1, 22) = 9.91$, $MS_e = .23$ $p < .01$. Masculinized faces were less well recognized ($d' = .76$) than unaltered ones ($d' = 1.07$). Again, in combination, these two ANOVAs indicate that the gender manipulation in and of itself did not affect recognition accuracy.

4.3. ***Discussion***

Thus, in both versions of the experiment, with image distortion statistically controlled, recognition memory was no better for unaltered faces than for faces with either same-gender or cross-gender alterations. Additionally, in both versions, there was no difference in the recognizability of cross- versus within-gender manipulations, indicating that faces more difficult to classify by gender were as easy to recognize as faces less difficult to classify by gender. Applying the same logic as Bruce et al. [7], these results are supportive of their model of face recognition: Gender and familiarity processing proceed in parallel.

Not only have we replicated Bruce et al.'s principal finding but we also have extended it. Strictly speaking, their result holds only for male faces in that they analyzed data only from their sample of male faces. Our findings were equally true of both face genders. In their study, the two degrees of masculinity of face, high and low, were determined by psychometric criteria applied to rating scale data–for instance, low rating scale values indicated male faces which had female characteristics. High rating scale values indicated highly masculine male faces. In our experiments, on the other hand, we manipulated a priori the degree of maleness or femaleness of each facial image. Hence, we could compare performance taken across unaltered face images at all levels of masculinity/femininity and across the same faces altered in a more masculine and then in a more feminine direction. Bruce et al. were only able to compare highly masculine (roughly equivalent to our masculinized males) and low masculine faces (presumably rather similar to our feminized male faces).

Bruce et al. [7] called for further research to identify sources of shared and unshared information for gender and familiarity judgments. Both in the present experiments and in prior ones [15, 16], we have shown that weights on the second and third eigenvectors (extracted from the cross-product matrix of pixel values associated with learned face images) are good predictors of face gender. The second and third eigenvectors are also within one of two bands of eigenvectors that provide useful information for recognition memory judgments [16, 19].

In Experiment 3, we have demonstrated that image degradation due to gender-related weight changes on the second and third eigenvectors negatively impacts recognition memory and statistically controlling for the effects of this degradation eliminates the decrease in recognition memory. In contrast, neither experimental nor statistical control of image degradation imparted by our gender manipulations eliminates their effects, either on gender categorization times and errors or on masculinity/femininity ratings (Experiments 1 and 2). Possibly the information for familiarity judgments contained in the second and third eigenvectors is more susceptible to image distortion/degradation than is the information for gender judgements. The information for familiarity judgments may be more "fragile," less "robust" than is the information for gender judgements. Alternatively, it may be that the image degradation resulting from gender manipulation changes the general distinctiveness of the altered faces. Perhaps due to their being distorted in the same way, all gender-altered faces might be more similar to one another than the unaltered faces and hence harder to recognize. In either case, cross-gender manipulations make face gender classification more difficult, but not face recognition, when the effects of image distortion are partialed out.

There is also a source of information for recognition/familiarity that is not shared with information for gender categorization. As O'Toole et al. [16] have discovered, a broad band of eigenvectors with relatively small weights provides even more useful information for recognition memory judgments than does the band that contains the second and third eigenvectors. Thus, when the uncontrolled image distortion induced in the faces of Experiment 3 resulted in a decrement in recognition memory performance of about one- third, quite possibly performance did not deteriorate further because the information for recognition contained in the eigenvectors with smaller weights was not similarly distorted. Certainly these latter eigenvectors had not had their weights adjusted in any way. It is also possible that image distortion induced by an increase in contrast in the second and third eigenvectors masked some of the information for recognition found in the higher-order eigenvectors, thereby producing at least some of the recognition decrement noted. Again, however, partialing out the effects of distortion induced by gender manipulation leaves the recognition process unaffected.

5. **General Discussion**

Combined, the results of Experiments 1 and 2 confirm the psychological relevance of the statistical features (macrofeatures, visually) derived from our computational model. We have shown that model-predicted sources of gender information may indeed impact gender categorization and ratings of femininity-masculinity by human observers. Cross-gender weightings on the second and third eigenvectors decrease both categorization speed and accuracy, as compared with unaltered faces and faces with increased same-gender weightings. Cross-gender weightings also produce the expected effects on gender ratings, with feminized male faces rated as more feminine than unaltered male faces and masculinized female faces rated as more masculine than unaltered female faces.

As previously noted, the macrofeatures constituting the second and third eigenvectors capture information for gender-relevant differences, in hairstyle and global face shape, among others. Their advantages over a priori selected task-relevant features are at least two. First, they are the statistical features from which face images may be constructed by appropriate weighted combination. This means that the pattern of weights on selected subsets of eigenvectors may well predict certain facial characteristics, gender, race, and attractiveness, for instance [cf. 15, 19]. Second, as was done in the present study, these eigenvector (macrofeature) weights may be manipulated in quantitative fashion to produce facial images whose psychological significance may be assessed by human observers.

Finally, we have found results not inconsistent with those of Bruce et al.'s [7] conclusion: The processes of face gender categorization and recognition (familiarity) are arguably independent. Bruce et al. proposed a model in which gender categorization and recognition processing proceed in parallel, with gender judgments completed first. We cannot speak to the latter aspect of their model, but for our observers as well as theirs, faces whose gender was more difficult to categorize (faces with cross-gender characteristics) were not more difficult to recognize.

We should, nevertheless, issue a caveat. Even though we have been able to partial out the visual "noisiness" of our gender manipulations, it would be desirable to confirm our findings with further studies wherein the information for gender categorization is manipulated without image degradation. Possible ways to reduce this noisiness would include morphing face images to a standard configuration [e.g., 14] and the work of Vetter and Poggio [26] aimed at solving a pixel-based correspondence problem for face images.

6. Acknowledgement

This project was supported by NIMH grant MH51765 awarded to Alice O'Toole.

References

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

[2] Anderson J A and Mozer M C, Categorization and selective neurons. In *Parallel Models of Associative Memory*, ed. by Hinton G E and Anderson J A (Erlbaum, Hillsdale, NJ, 1981) pp. 213–236.

[3] Anderson N H, Scales and statistics: Parametric and nonparametric. *Psychol Bull* **58** (1961) 305–316.

[4] Benson P J and Perrett D I, Face to face with the perfect image. *New Scientist* **1809** (1992) 32–35.

[5] Brown E and Perrett D I, What gives a face its gender? *Perception* **22** (1993) 829–840.
[6] Bruce V, Burton A M, 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.
[7] Bruce V, Ellis H D, Gibling F and Young A W, Parallel processing of the sex and familiarity of faces. *Canad J Psychol* **41** (1987) 510–520.
[8] Bruce V and Young A W, Understanding face recognition. *Brit J Psychol* **77** (1986) 305–327.
[9] Burton A M, Bruce, V and Dench N, What's the difference between men and women? Evidence from facial measurement. *Perception* **22** (1993) 153–176.
[10] Craw I and Cameron P, Parameterising images for recognition and reconstruction. In *Proc Brit Machine Vision Conf*, ed. by Mowforth P (Springer Verlag, London, 1991) pp. 367–370.
[11] Ellis H D, Shepherd J W and Davies G M, Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception* **8** (1979) 431–439.
[12] Ellis H D, Theoretical aspects of face recognition. In *Perceiving and Remembering Faces*, ed. by Davies G, Ellis H and Shepherd J (Academic Press, London, 1981) pp. 171–197.
[13] Ellis H D, Processes underlying face recognition. In *The Neuropsychology of Face Perception and Facial Expression*, ed. by Bruyer R (Erlbaum, Hillsdale, NJ, 1986) pp. 1–27.
[14] Hancock P J B, Burton A M and Bruce V, Face processing: Human perception and principal components analysis. *Mem and Cog* **24** (1996) 26–40.
[15] O'Toole A J, Abdi H, Deffenbacher K A and Bartlett J C, Classifying faces by race and sex using an autoassociative memory trained for recognition. In *Proc 13th Ann Conf Cog Sci Soc*, ed. by Hammond K J and Gentner D (Erlbaum, Hillsdale, NJ, 1991) pp.847–851.
[16] O'Toole A J, Abdi H, Deffenbacher K A and Valentin D, Low- dimensional representation of faces in higher dimensions of the face space. *J Optical Soc Amer* **10** (1993) 405–411.
[17] O'Toole A J, Deffenbacher K A, Abdi H and Bartlett J C, Simulating the "other-race effect" as a problem in perceptual learning. *Connect Sci* **3** (1991) 163–178.
[18] O'Toole A J, Deffenbacher K A, Valentin D and Abdi H, Structural aspects of face recognition and the other-race effect. *Mem and Cog* **22** (1994) 208–224.
[19] O'Toole A J, Deffenbacher K A, Valentin D, McKee K, Huff D and Abdi H, The perception of face gender: The role of stimulus structure in recognition and classification. *Mem and Cog* in press.
[20] O'Toole A J and Thompson J L, An X Windows tool for synthesizing face images from eigenvectors. *Behav Res Meth, Instr, and Computers* **25** (1993) 41—47.
[21] O'Toole A J, Vetter T, Troje N F and Blthoff H H, Sex classification is better with three-dimensional head structure than with image intensity information. *Perception* **26** 26–75.

[22] Poggio T and Beymer D, Learning networks for face analysis and synthesis. In *Proc Int Workshop on Face and Gesture Recognition*, ed. by Bichsel E M (Zurich, Univ of Zurich Multimedia Laboratory, 1995) pp. 160–165.
[23] Roberts T and Bruce V, Feature saliency in judging the sex and familiarity of faces. *Perception* **17** (1988) 475–481.
[24] Sheperd J W, Social factors in face recognition. In *Perceiving and Remembering Faces*, ed. by Davies G M, Ellis H D and Sheperd, J W (Academic Press, London, 1981) pp. 55–79.
[25] Valentin D, Abdi H, O'Toole A J and Cottrell G W, Connectionist models of face processing: A survey. *Pattern Recog* **27** (1994) 1208–1230.
[26] Vetter T and Poggio T, Image synthesis from a single example image. In *Proc European Conf on Computer Vision* (Cambridge University Press, Cambridge, U.K., 1996) pp. 652–659.
[27] Yamaguchi M K, Hirukawa T and Kanazawa S, Judgment of gender through facial parts. *Perception* **24** (1995) 563–575.