



Technical Report No. 094

**Categorical perception of
gender:
No evidence for unfamiliar faces**

Isabelle Bülthoff,¹ Fiona N. Newell,²

October 2005

¹Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany,
email: isabelle.buelthoff@tuebingen.mpg.de

²Trinity College, Dublin, Ireland

Categorical perception of gender: No evidence for unfamiliar faces

Isabelle Bühlhoff and Fiona N. Newell

Abstract. We investigated whether male and female faces are discrete categories at the perceptual level. We created artificial gender continua between male and female faces using a 3D-morphing algorithm and used classical categorization and discrimination tasks to investigate categorical perception of gender. In Experiments 1 and 3, 3D morphs were computed between male and female faces. The results of the discrimination task suggest that the gender of unfamiliar faces is not categorically perceived. When participants were familiarized with the male and female endpoint faces before testing (Experiment 3), a categorical effect was found. In Experiment 2, only shape or texture of unfamiliar 3D morphs was indicative of gender, while other information (e.g. texture or shape) was kept constant. Again there was no evidence of a categorical effect in the discrimination task. In Experiments 1, 2 and 3, changes in the gender of a face were also coupled with changes in identity which may have confounded the findings. In Experiments 4 and 5, we used face continua in which only the gender of the facial features changed, while the characteristic of the facial features remained constant. When the faces were unfamiliar (Experiment 4), there was no evidence of categorical perception of gender. In Experiment 5, participants learned to classify the face images in two gender categories using a feedback procedure. A clear categorical effect for gender was present after training. Our findings suggest that despite the importance of faces, gender information present in faces is not naturally perceived categorically. Consequently participants showed categorical perception of gender only after training with the face stimulus set.

1 Introduction

Usually we have little trouble in determining the gender of a person that we meet for the first time. Nevertheless, most of us have experienced the complexity of this seemingly easy task when confronted by a young person in baggy clothing, leading us to scrutinize his/her face, body, voice and body language for any telltale signs of his/her gender. Here we will be concerned with how we perform this task when the only information available is static visual information given by the shape and the texture¹ of an adult face. More specifically, we will examine if we automatically perceive a face as male or female, i. e. if we represent faces as two discrete categories, one for female faces and one for male faces.

In our visual world we are confronted with a multitude of objects, many of which are moving and changing. One strategy to simplify the mental representation of the external world is to organize all stimuli into discrete categories (e.g. Rosch et al. [1976]). While some categories are physically very different from each other, (for examples the categories of cars and insects) others are very similar (for examples the categories of happy and sad faces). The visual system has a clever way of discriminating between highly similar items in our world by exaggerating small perceptual differences, thus creating clear boundaries between groups of items. This phenomenon, i.e., the fact that we perceive a range of monotonically varying stimuli as belonging perceptually to different categories, is called categorical perception (Harnad [1987]). Objects within a category are perceived as more similar to each other than to objects belonging to another category even if the physical differences between them are equal. One well-known example of categorical perception is our perception of a rainbow. Although a rainbow constitutes a monotonic increase of wavelength of visible light, we perceive a series of discrete color bands. Furthermore, a wavelength in the yellow band of the rainbow is perceived as qualitatively more similar to another yellow wavelength than to a wavelength of the same physical distance but belonging to the orange band of the rainbow.

Many instances of categorical perception (CP) have been described in the literature. CP was first observed with auditory stimuli (Burns and Ward [1978]; Eimas et al.; Liberman et al. [1967]; amongst many others) and color

¹Also called pigmentation or skin color of the face.

(for example Bornstein [1987]; Bornstein and Korda [1984]; Boynton [1979]; Valois and Valois), but it has since been found in a variety of domains. Recently, complex visual stimuli have also been shown to be perceived categorically. In particular, CP has been reported for facial expressions (Calder et al. [1996]; de Gelder et al. [1997]; Young et al. [1997]) and for facial identity (Beale and Keil [1995]; Levin and Beale [2000]). There are a few studies demonstrating CP for complex visual stimuli other than faces, for example, histological slides (Adamson and Sowden [2000], Adamson and Sowden [2001]) and man-made familiar objects (Newell and Bülthoff [2002]).

Male and female faces are well known categories; generally we do not need to know a person to be able to identify the gender of their face on a photograph (Bruce et al. [1993], but see Bruce [1986], for the role of familiarity on gender decisions). The same is true for facial expressions (Ekman [1994]). During infancy, children learn to categorize humans into male and female categories not by viewing faces alone, but with the help of a multitude of other cues. Later, adults and older children can determine with a high degree of accuracy the gender of a person by viewing an image of the face only (Bruce et al. [1993]; Wild et al. [2000]). The question we want to address here is how we perceive facial information pertaining to the gender of a person. We investigated whether gender discrimination is mediated by perceptual categorization as was found for facial expressions. To study categorical perception of, for example, facial expressions, one typically creates morph sequences between two images of the same face showing different facial expressions. The resulting face stimuli from the morph sequence are then used for categorization and discrimination tasks. While the existence of continuous changes between different expressions seems like a naturally occurring phenomenon within the same individual, this is clearly not the case for gender. However, variation from the gender prototypes can exist at the population level. Furthermore, gender information is not confined to the face because body shape is also an important cue to the gender of a person. Nevertheless we excel in tasks pertaining to the categorization of gender and to the categorization of facial expressions.

With new media technology and computational methods (Blanz and Vetter [1999]), it is possible to create artificial gender continua between three-dimensional computer-reconstructed male and female faces, thus creating face images ranging along a gender continuum. We created and used such stimuli in our experiments to investigate how gender in faces is perceived. Gender perception from face information alone may not be obvious because in our every day life we might use more reliable cues such as body shape, and cultural cues such as jewelry and hair length.

We followed the classical procedure for testing for categorical perception. In this procedure, participants are first required to discriminate between pairs of face images that differ by equal physical increments between two different faces in a discrimination task. Thereafter participants are required to label each face image as one of two pre-specified categories in a categorization task. A CP effect occurs when (1) participants classify most faces on one side of the gender boundary defined in the categorization task as belonging to one category and the other faces as belonging to the other category with a sharp category boundary, and (2) participants can more easily discriminate between pairs of stimuli that straddle that boundary than between pairs that are entirely within one category or the other. In other words, in CP the category boundary defined by the categorization curve obtained in the classification task predicts the peak in performance in the discrimination task.

A summary of our predictions is as follows: If gender perception for faces is categorical, we predicted that in the categorization task all faces would be perceived as male or female, with a sharp change at the subjective category boundary even though all faces presented are evenly distributed along the artificial gender continuum. More importantly, in the discrimination task we expected pairs of faces to be discriminated more accurately when they straddled the category boundary estimated from the results of the categorization data than when both faces belonged to the same category.

2 General Method

Stimuli Database. We used a database of 100 female and 100 male faces collected by Vetter, Troje and coworkers (O’Toole et al. [1997]; Troje and Bülthoff [1996]) to generate our stimuli. These faces were derived from three-dimensional scans of male and female participants (18-45 years old) wearing swimming caps obtained using a laser scanner (CyberwareTM 3030PS). All faces were devoid of secondary cues which could connote the gender of the face such as makeup, accessories or facial hair. Most faces in the database had previously been rated for distinctiveness (O’Toole et al. [1998]). We picked 6 male and 6 female faces with similar, low distinctiveness

ratings that were most suitable for the morphing procedures. In all experiments we used face stimuli derived from some or all of these 12 faces except in one experiment (Experiment 4a) where average faces derived from the whole database were used.

Scanning method. For each participant, the scanner created a profile of his/her face by shining a low-intensity laser beam in the shape of a vertical stripe onto the head. The laser beam moved around the head in 15s while a video sensor captured the profile at a rate of about 30 times per second, sampling the shape of the head on a regular cylindrical grid of 512 X 512 points with a resolution of 0.8 degrees horizontally and 0.615 mm vertically. Simultaneously, a second video sensor acquired color information in the same spatial resolution. Consequently each scan is represented by two sets of data of 512 X 512 points each. One set describes the 3D shape of the head (geometric data), the other one the RGB-color values of each pixel of the head image (textural data). The swimming caps and the shoulders were subsequently removed as part of the post-processing of the scan data. The resulting faces were devoid of hair or scalp, but included the ears and ended at the neck. After processing, each face was represented by approximately 7×10^4 vertices and the same number of color values. A normalization routine placed each face at a standard orientation and position in space.

Morphing method. For our study we generated new faces based on all or some of the laser-scanned faces. We used a correspondence method developed by Vetter, Blanz and colleagues (for more details see Blanz [2000]; Blanz and Vetter [1999]; Vetter and Poggio [1997]). With this method, based on optic flow algorithms, corresponding vertices between laser-scanned faces can be found automatically. Thus all faces of the database are in dense point-to-point correspondence with each other. With this powerful method it is possible to create new faces by forming linear combinations of faces. We used this method to compute 3D-morphs between pairs of laser-scanned faces (see Experiment 1 and 3) and to generate variations of an average face calculated from all faces present in the database (Experiment 4a). We also used this method to calculate gender variations of the facial features of individual faces (Experiment 4b). Furthermore, as texture data and geometric data of a face can be computed independently of each other, it is possible to superimpose any facial texture onto any face shape, meaning that the texture of one face can be applied onto the shape of another face (Experiment 2).

Gender continua. We created gender continua based on the faces from the database to investigate how gender in faces is perceived. Each continuum consisted of a set of eleven face images. In each set there were nine morphs evenly distributed in 10% increments between the endpoint faces. Endpoint face image 0 was 0% male and 100% female; face morph image 1 was 10% female and 90% female, etc. with endpoint face image 10 being 100% male and 0% female. The different types of gender continua will be described in more detail within the appropriate experimental sections.

In the present study, all faces were presented on a black background either frontally (i.e. full-face) or rotated 20° to the right from the observer's point of view. The rotated faces were more informative about the 3D shape of the face, especially the size and shape of the nose and the jaw, (see Bruce et al. [1993] and Burton et al. [1993]).

Apparatus. The digital face images (256 x 256 pixels in size and 8 bit per color channel) subtended approximately 6° x 6° of visual angle. The average viewing distance was 57 cm. All experiments were conducted in a darkened room. For the Experiments 1, 2 and 3, the stimuli were presented using Psyscope (Cohen et al. [1993]) on a Power Macintosh 9500/132 linked to a standard color monitor. In Experiments 4 and 5 we used Eprime software (Psychological Software Tools, Inc.) on a personal computer running Windows 98. The experiments were run in a darkened room. A button box was used for collecting responses.

Procedure. Before performing any task, participants were required to read the instructions on the screen, and to hit a button of a button box to indicate their readiness to start the task. In all experimental sessions participants performed a discrimination task followed by a categorization task with a self-timed break between the two tasks. These specific tasks will be described in more details below. The subjective category boundary for each face combination was estimated from the results of the categorization task. This determined which face pairs straddled the category boundary in the discrimination task. Because the categories were named and their boundary defined in the categorization task, this task always followed the discrimination task in order to avoid biasing the

discriminations. A short practice block preceded the test blocks in the discrimination task.

Participants. All observers (age range 18 to 42 years) were paid volunteers and naive as to the purpose of the experiments. Each participant took part in one experiment only and did not participate in any of the other experiments. Most of the participants were undergraduate students from the Eberhard-Karls University of Tübingen, Germany. All participants had normal or corrected-to-normal vision.

3 Experiment 1

In this experiment, our purpose was to investigate whether gender continua created between unfamiliar male and female faces were perceived categorically.

3.1 Method

Participants. There were ten participants in the following experiment. Four participants were female².

Stimuli. Gender continua were created between each of all possible pairs of 6 male and 6 female laser-scanned faces, i.e. for 36 continua. Each continuum consisted of the original male and female endpoint faces and nine equidistant morphs. Both shape and texture of the face stimuli were informative about the gender of the face. All faces and morphs were presented rotated 20° to the right. For examples of the face stimuli used see Figure 1

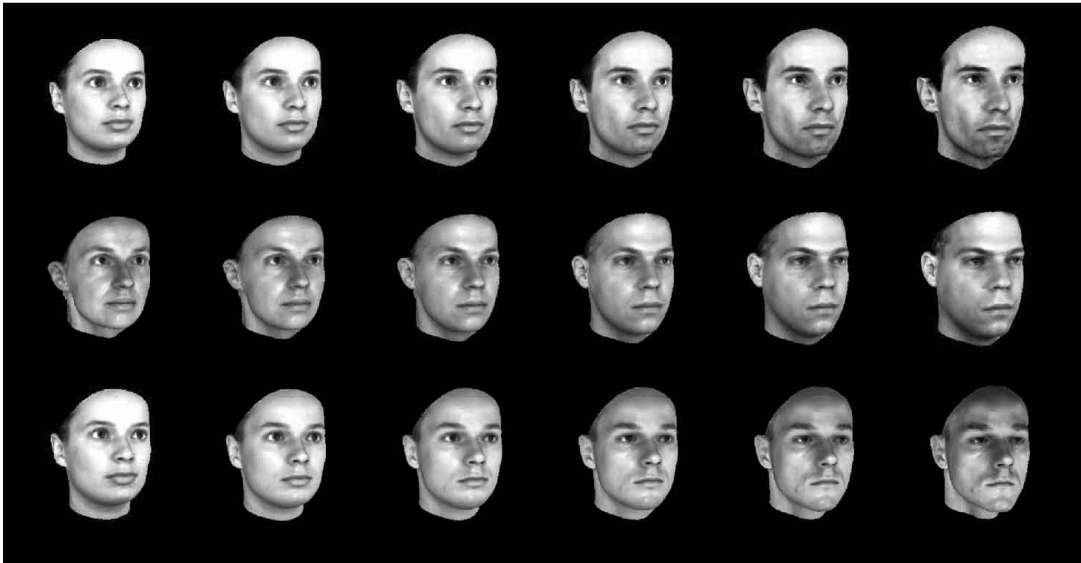


Figure 1: Illustration of three of the 36 gender continua used in Experiment 1 and two of the 6 gender continua used in Experiment 3 (upper and middle row). Note that the same female face is used twice (upper and lower row). Alternate face images from each pictured continua is shown. The endpoint face images on the very left and very right correspond to the original faces.

Design. The experiment was divided into two separate tasks; a discrimination task and a categorization task. The *discrimination task* was based on an XAB match-to-sample design used by Newell and Bühlhoff [2002] (see also Calder et al. [1996]; Etcoff and Magee [1992]; Young et al. [1997]) in which an image of a face (stimulus X) was presented initially in a trial followed by two face images (stimuli A and B) presented simultaneously, left and right of fixation. Stimuli A and B (image pair) were always physically different from each other and stimulus X was identical to either stimulus A or B. Stimuli A and B differed by 2 morph steps (e.g., face images 1 and 3) along the gender continuum. All stimuli in one trial belonged to the same gender continuum and their order of presentation was counter-balanced. The *categorization task* was a forced choice paradigm. Participants were shown all faces

²In pilot experiments, we did not find any difference in performance between male and female participants.

of all continua one by one in random order and were asked to classify each face image as male or female.

There were 648 trials (36 gender continua, 9 image pairs, repeated twice) in the discrimination task and 336 trials in the categorization task (36 gender continua, 9 morphs images shown once and 12 endpoint images shown once). All trials were randomly presented within each task on a participant by participant basis.

Procedure. In the discrimination task, there were three face stimuli shown in any one trial. A fixation cross preceded the face stimuli for 250ms in each trial. The first face image (X) was shown for 750 ms in the centre of the screen followed by a blank screen for 1000ms. The next pair of stimuli (A and B) remained on the screen until the participant pressed a response button. Each of the A and B stimuli were displayed 3cm to the left and right of the centre point of the screen. An inter-trial interval of 500ms followed the participant's response. In order to acquaint participants with the XAB procedure, the experiment began with a random selection of 8 practice trials. The training block was followed by three experimental blocks and participants received a self-timed break between blocks. Participants were instructed to respond as fast and as accurately as possible, indicating which face image of the AB pair was identical to the preceding face image X. Participants were instructed to press the left (or right) response button of the button box to indicate that the left image (or right image) was identical to the first image. In the categorization task, each trial began with a 500ms fixation cross. A face image then appeared and remained on the screen until the participant responded. An inter-trial interval of 500ms followed each participant's response. Participants were instructed to decide as fast and as accurately as possible whether each face image shown was male or female. The right response button of the button box was assigned as the "female" response key, and the left was assigned the "male" response key. The face images in the categorization task were always the same as those shown in the discrimination task. Participants took approximately 90 minutes to complete both tasks of the experiment.

3.2 Results

The mean number of correct responses made to the XAB task and the mean frequencies with which participants categorized each image as female in the gender categorization task are shown in Figures 2 (upper and middle row).

Categorization task The subjective category boundary was determined as the point at which the categorization function crosses the 50 gender response level for all face combinations. Figure 2 (upper row) shows that the point of subjective category boundary lay between face image numbers 3 and 4.

Discrimination task Performance was calculated across all participants and all face combinations (Figure 2 (middle row)). On average, performance was 62.3% correct. We conducted a paired t-test using category position as the factor. The category position factor refers to the discrimination performance at pairs of face images that lay at either end of the gender continua (i.e. the average performance to face image pairs 0-2 and 8-10) and the discrimination performance at image pairs that straddle the category boundary (i.e. faces images 4-6). The mean percentage of correct responses made to end-pair images was 62.8% and for the images straddling the category boundary it was 65.1%. We found no effect of category position for either subjects [$t(9) = -0.638$, n.s.] or items [$t(35) = -1.559$, n.s.].

Reaction times. The reaction time across all trials in the discrimination task was 1793 ms. The average reaction times for each face pair are plotted in Figure 2 (lower row). We conducted a one-way ANOVA on the average reaction times to the different face image pairs in the discrimination task. We found a significant effect of subjects [$F(8,72) = 3.354$, $p < 0.01$] and of items [$F(2(8, 280) = 4.300$, $p < 0.001$]. A post-hoc Newman Keuls analysis revealed that the reaction times to the image pair 0-2 were significantly slower than RTs to image pairs 4-6, 6-8 and 7-9 ($p < 0.05$). Reaction times to the pair of images straddling the category boundary (3-5) were therefore not significantly faster than to pairs of images lying within each category.

3.3 Discussion

Categorization task. The categorization plot indicates that the subjective gender boundary lies between faces 3 and 4, although in physical image-based terms the boundary lies at face 5 which is equally male and female. This difference denotes a male bias in the categorization judgment of the participants. Likewise, categorization performance for the endpoint faces (faces 0 and 10) shows that on average 23.2% of the female endpoint faces

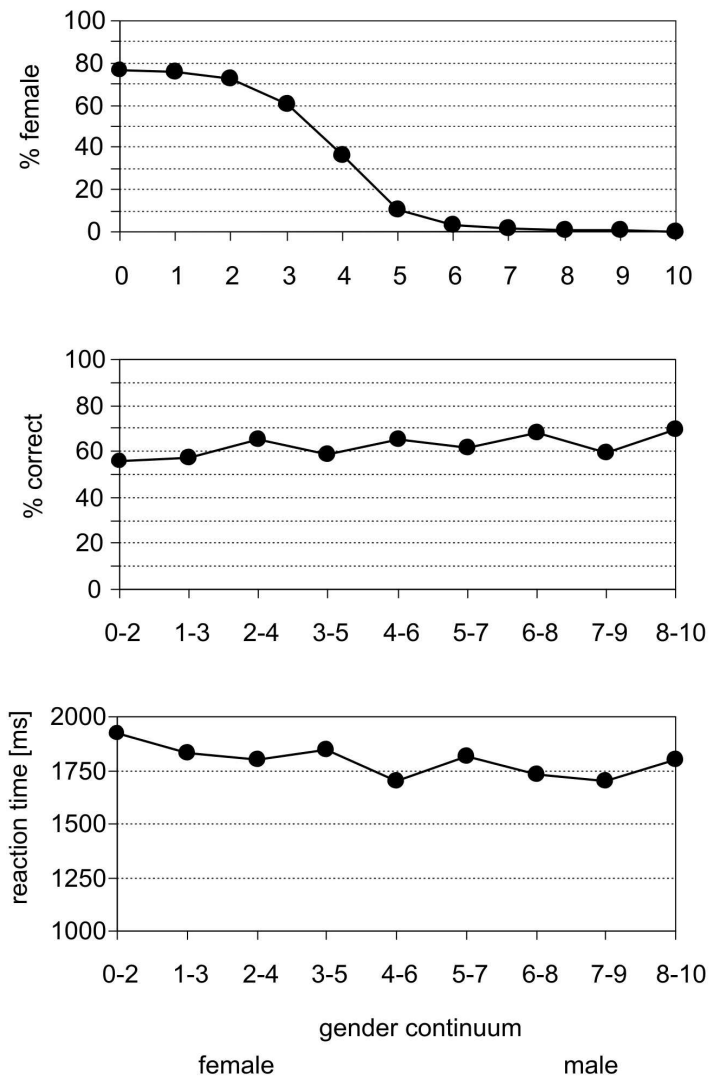


Figure 2: Plots showing data from Experiment 1: (upper row) mean categorization data, (middle row) mean discrimination data and (lower row) mean reaction times in the discrimination task. The categorization plot represents the mean percentage female responses to each face stimulus along the continuum. The discrimination plot represents the mean percentage of correct responses to each AB pair. In the upper row the gender continuum is represented by individual faces, in the middle and lower row by face pairs.

(image number 0) were misjudged as male faces while male faces (image number 10) were never misjudged as female. Bruce et al. (1993) in their study with laser scans of faces have also reported that the gender of female faces is more often misjudged than that of male faces. Other researchers who used images (not laser scans) of adults' or children's' faces have also found a male bias in gender judgment (Cheng et al. [2001]; Intons-Peterson [1988]; Wild et al. [2000]). Potential reasons for a male bias are numerous. There could be social as well as perceptual reasons for this male bias. Since we used faces only rather than the entire head, the judgment of female faces might be affected by the absence of hair. Although adults know that the sex of a person does not strictly correlate with presence and/or length of hair and makeup (and so on), it might be hard for them to ignore the absence of hair in our face stimuli. Another, perhaps more significant, factor appears to be that there are no positive markers for female adult faces, while there are some for male faces such as the presence of facial hair (or stubble), bushy eyebrows, strong jaws, etc.. Female faces in contrast keep many features of a child's face, for example small, round faces, big eyes and smooth skin (e.g. Abdi et al. [1995]; Clemente [1985]; Zebrowitz [1997]). In our experiments, many participants reported categorizing female faces as male because they judged

these faces to be young male faces. Thus some female faces might have been mistaken for male because of some ambiguity about the age as well as the sex of the faces. Finally, another factor contributing to the male bias might be perceived attractiveness. In our database, most laser-scanned faces look less attractive than the original faces. We speculate that some observers might have categorized female faces as male because of their expectancy for female faces to be more attractive.

The gender information was varied continuously in the 9 faces between the two endpoint faces. Participants were able to extract gender information present in the faces and to use it for categorization. The sharp step-like function in the categorization data plotted in Figure 2 (upper row) indicates that category differences, not just proportion of gender mixture, are affecting categorizations and that all observers set their perceptual gender boundary at approximately the same place along the gender continuum for all face combinations. The presence of a sharp step indicates furthermore that, despite the male bias discussed above, participants perceived clearly two gender categories; an obvious prerequisite for the potential presence of a categorical effect. Whether perception is categorical or not is determined by the results of the discrimination task discussed below.

Discrimination task. A comparison of the within-category versus between-category discrimination performance was conducted across all participants and all items and did not reveal any categorical effects. Pairs of faces straddling the gender category were not more easily discriminated than pairs of faces belonging to the same gender category. We investigated results for each face combination and each participant separately to ensure that the averaging process did not mask discrimination performance for some face combinations. Scrutiny of the observers' performance for all face pairs separately shows that, even for face combinations for which the gender categorization task was easier (i.e. a steeper step), discrimination was not better or more categorical than for other faces. The results suggest that we do not perceive the gender of unfamiliar faces categorically. An analysis of the reaction times show a similar pattern to the accuracy data, thus we found no evidence of a speed-accuracy tradeoff; i.e. participants were not less accurate when they answered quickly.

Despite the fact that performance was over chance level, we were concerned that many of our participants complained about the difficulty of the discrimination task. To ensure that the absence of CP in Experiment 1 was not due to the fact that the discrimination task was too difficult for the observers, in a series of control experiments we repeated the same test with images which were 3 steps apart. Thus it should have been easier for the observers to choose the face identical to the X stimulus in the test pair as the distractor face differs more strongly from the X stimulus. Indeed we found that by increasing the image step size the overall correct response rate increased in the discrimination task although, we again failed to find evidence for categorical perception. We found 14.0% errors in total, 11.7% errors were to the end points and 12.8% errors to the image pairs across the boundary.

These findings suggest that the difficulty of the task in the present experiment did not obscure effects of CP which might otherwise have been present. Other possible reasons why we did not find CP effects for gender will be explored in the following experiments.

4 Experiment 2

Studies of Bruce and Langton [1994] have shown the importance of skin color for recognition of face gender. In the previous experiment, facial texture was not normalized and the facial skin tone varied for each face continuum. For some gender continua involving endpoint faces of markedly different skin tones, the overall texture information changed visibly from one face of the sequence to the next. Texture variations between these faces might have been so salient that the participants did not pay attention to the face itself, but tried to perform the categorization and the discrimination task exclusively on the texture of the whole face, thereby overlooking more subtle gender variations necessary for CP of gender. To test whether skin tone differences might have hindered CP of gender, we repeated the same experiment as above with another set of faces in which only shape was indicative of gender while the texture did not vary. For comparison we also created a set of faces that were free of shape variations, therefore, only texture was indicative of gender. In this way we could investigate directly the importance of face texture and face shape for gender discrimination.

4.1 Method

Participants. Twelve students participated in the following experiment. Eight of the participants were female.

Stimuli. We created two qualitatively different gender continua from each face combination; a shape-based continuum and a texture-based continuum. As mentioned in the general method section, the correspondence method allowed us to compute separately the texture and shape of faces and to recombine them with any other faces to create new faces. We computed separately the average texture and the average shape based on the database of 200 faces. *Shape-based gender continua:* The average texture was combined with the changing shapes of morphs between individual male and female laser-scanned faces. Only the shape was indicative of gender in these continua. *Texture-based gender continua:* The average shape was combined with the changing texture of morphs computed between individual male and female faces. Consequently, in these continua only texture was indicative of gender.

Note that the endpoint faces were not the original faces (as was the case in Experiment 1) but consisted of the original shape (shape-based continua) or texture (texture-based continua) of each laser-scanned face combined with the average texture or shape respectively. All faces were presented in full-face view. An illustration of both types of continua is shown in Figure 3. A selection of 4 female faces and 4 male faces was taken from the face set used in Experiment 1. All pairwise combinations were used in this experiment, thus there were 16 face combinations. Each of the 16 face combinations were rendered twice, once for the shape-based gender continua and once for the texture-based gender continua.



Figure 3: Sample set of the two types of gender continua used in Experiment 2. In the upper row an example of a texture-based continuum is shown. An example of a shape-based continuum is illustrated in the lower row. Every other face image of the pictured continua is shown. Note that in both types of continua the computed endpoint faces do not correspond to the original faces from our database. Instead the endpoint faces have either the shape or the texture of the neutral face. See text for more information.

Design and procedure. The experiment was based on a 2-way, within-subjects design with type of continuum (shape-based or texture-based) and image pair (discrimination task) or image number (categorization task) as factors. Participants took a self-timed break between blocks. The *discrimination task* began with 8 random practice trials. The experimental session consisted of two main blocks, the texture-based and shape-based blocks of images. The order of the blocks was counterbalanced across participants. The procedure was the same as in Experiment 1. There were 576 trials in the discrimination task (2 types of continua, 16 continua, 9 image pairs, repeated twice). A *categorization task* followed the discrimination task. The procedure was the same as in Experiment 1. All face stimuli were presented randomly in one block. There was a total of 352 trials in the categorization task (2 types of continua, 16 continua, 11 face images). Participants needed about 70 minutes to complete this experiment.

4.2 Results

The mean rate of female responses in the gender categorization task and the mean number of correct responses made in the XAB tasks for texture-based and shape-based gender are shown in Figures 4 (upper and middle row).

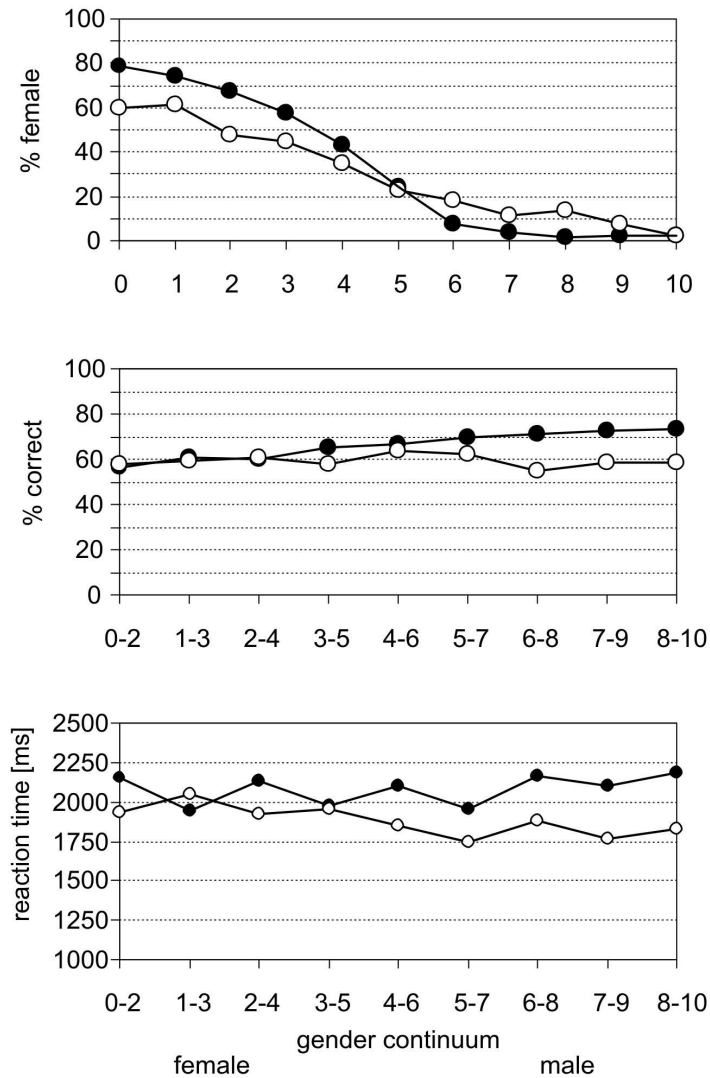


Figure 4: Plots showing data from Experiment 2: (upper row) mean categorization data (middle row) mean discrimination data, (lower row) mean reaction times in the discrimination task. The curves with filled symbols show data for all texture-based gender continua. The curves with unfilled symbols show data for all shape-based gender continua.

Categorization task. The subjective category boundary was determined by the categorization performance for each of the 16 continua. As in the previous experiment, we determined the category boundary as the point at which the categorization curve crosses the 50% gender level for each face combination. The category boundary lay between face images 3-4 for the face continuum with texture changes and close to face image 2 for the face continuum involving shape changes only.

Discrimination task. Overall performance was around 66 % correct for texture-based continua and 59% correct for shape-based continua. We conducted a two-way, repeated measures ANOVA using category position and type of continua (shape or texture) as factors. As in Experiment 1, the category position factor refers to the discrimination performance at pairs of face images that lay at either end of the gender continua and the

discrimination performance at image pairs that straddle the category boundary. For the shape continuum, the mean number of correct responses made to end-pair images was 58.1% and for the images straddling the category boundary it was 60.7%. For the texture continuum, the mean number of correct responses made to end-pair images was 65.0% and for the images straddling the category boundary it was 66.7%. We found no effect of category position for either subjects [$F(1,11) = 1.073$, n.s.] or items [$F(2,15) = 1.359$, n.s.]. We found a main effect of type of face continuum for subjects [$F(1,11) = 17.817$, $p < 0.01$] and for items [$F(2,15) = 5.821$, $p < 0.05$]. Thus performance was significantly better for texture-based than for shape-based continua but there was no CP for either type of continua. There was no interaction between the factors, [$F(1,11) = 0.084$, n.s.; $F(2,15) = 0.044$, n.s.].

Reaction times. The mean reaction time for the texture continua was 1881 ms, and 2079 ms for the shape continua. The mean reaction times to the discrimination tasks are plotted in Figures 4 (lower row). We conducted a one-way ANOVA on the mean reaction times to the different face image pairs in the discrimination task. We found no effect of face image pairs [$F(8, 88) = 1.513$, n.s.; $F(8, 120) = 1.699$, n.s.]. An effect of type of face continuum was almost approaching significance for subjects [$F(1,11) = 4.204$, $p = 0.065$] and was significant for items [$F(1,15) = 33.760$, $p < 0.001$]. There was an interaction between the factors for subjects analysis [$F(8,88) = 2.352$, $p < 0.05$], and approaching significance for the items analysis [$F(8,120) = 1.956$, $p = 0.0575$]. The interaction was probably due to the difference in the reaction time functions in discriminating face pairs across the shape and texture continua. For example, from Figure 4 (lower row) we can see that reaction times were faster to the male end of the texture-based continua than the male end of the shape-based continua of faces.

4.3 Discussion

Categorization task For texture-based continua, face images 0 were judged to be female in 79% of the trials whereas face image 10 were judged to be male in 100% of the trials. These results and the shape of the categorization curve in Figure 4 (upper row) are similar to the categorization results when shape and texture of the face stimuli were informative (Figure 2 (upper row)). For shape-based continua, the corresponding percentages were 60% and 98% and the shape of the categorization curve is flatter. Overall, these results indicate that texture information was more informative than shape for gender classification when faces were viewed frontally. Our finding that gender categorization accuracy was better with texture information alone than with shape information alone is in accordance with the findings of Bruce et al. (1993). In Experiment 2 of their study, these authors used texture-free laser scans and regular images of male and female faces as stimuli. They compared participants' performance on a gender decision task when both textural and shape information or only shape information was present. In the absence of textural cues, female faces were much more often misjudged (52.1% correct) than male (97.9%), while performance was 93.8% correct for female faces and 98.2% correct for male faces when textural and shape information was combined. Furthermore, they found that textural information is predominant in frontal images where 3D information is poor, i.e. where facial shape (especially nose and chin) is less prevalent. Our results are also in agreement with the studies of Hill et al. [1995] who compared the importance of texture and shape of faces in judgments of gender and race. They found that participants were more accurate in their judgments of gender when they were presented with textural data only than with shape-only laser-scans of faces.

As in Experiment 1, we found a male bias for both types of continua, i.e. participants showed a tendency to judge more face stimuli along a continuum to be male than female. This male bias was especially apparent in the shape-based continua for which a neutral texture was used. Note that Bruce et al. [1993] found an even stronger male bias in their experiment when they used texture-free face stimuli.

Discrimination task. Generally participants were faster and better in their decision when viewing face pairs of the texture-based continua than face pairs of the shape-based continua. We found no indication of a categorical effect in either accuracy performance or in response times when shape or texture alone was the indicator of the gender of the face. More specifically, when participants had to rely on shape information because there were no texture variations between faces, their overall performance dropped as might be expected, but, again, no categorical effect appeared. This suggests that the textural variations between face stimuli in Experiment 1 did not hinder any potential effect of CP of gender.

5 Experiment 3

It has been suggested that familiarity with the endpoint faces on a continuum is a prerequisite for CP of facial identity or expression (Beale and Keil [1995]; Etcoff and Magee [1992]). Recent studies also show that CP of identity is also present after short term learning of previously unfamiliar faces (Levin and Beale [2000]). In the following experiment, we tested whether familiarization with the endpoint faces would facilitate CP for face gender. Familiarity was manipulated by exposing the participants to the endpoint face stimuli prior to the experiment.

5.1 Method

Participants. Twenty participants took part in the following experiment. Ten of the participants were female.

Stimuli. The same six male and six female faces used in Experiment 1 were used in this experiment. However, here each face was paired only once, giving a total of six gender continua. As in the previous experiments, 11, equally distant image steps were extracted from each continuum.

Design. The experiment was based on a within subject design. In this experiment, the images from each of the six continua were tested in separate blocks. Our reason for testing face images from a single face pair combination in the same block was twofold. First, we wanted to reduce the difficulty of the discrimination task, as all participants in Experiment 1 found the discrimination task quite demanding. Second, we wanted to facilitate potential effects of CP. In the discrimination task, each image pair was repeated 6 times, giving a total of 324 trials (6 face continua, 9 image pairs, repeated 6 times). As in the previous experiments, one practice block of 8 randomly chosen trials preceded the six experimental blocks in the discrimination task. In the categorization task, each image was shown twice giving a total of 132 trials (6 gender continua, 11 images, repeated twice).

Procedure. Before performing the experiments, all participants were familiarized with the 6 male and 6 female original faces for 10 minutes. The images of all faces shown from a 20-degree view were displayed together on two pages (one for each sex). We labeled each face with a short first name, and each name clearly connoted the gender of the face, e.g. Kurt and Heidi (at least for our German participants the gender was clear from these names). Participants were told to view carefully the face images and to learn their names because they would be asked to recognize those particular faces later. In the subsequent recognition test the participant had to name each randomly presented face image. We set a performance criterion of 11 out of a possible 12 correct responses in the recognition task before the participant could continue with the experiment. Most of the participants reached this criterion after one learning block. The learning block was repeated for those who did not reach criterion. After successful completion of the recognition task, participants were told that the names would never be used in the subsequent experiments and proceeded to perform a discrimination task and a categorization task following the procedure described in Experiment 1. Participants needed about 50 minutes to complete this experiment.

5.2 Results

The mean number of correct responses made to the XAB task and the mean percentage of female responses to the gender categorization task are shown in Figures 5 (upper and middle row).

Categorization task. The subjective category boundary was determined by the categorization performance as described in the previous experiments. In this experiment, the category boundary lay between face images 4-5.

Discrimination task. The average correct performance over all face pairs was 65%. We conducted a paired t-test using category position as the factor (i.e. comparing the average performance to face image pairs 0-2 and 8-10 to performance for face images 3-5). The mean number of correct responses for the end pairs of face images was 61.7% and 69.4% for the images straddling the category boundary. We found a significant effect of category position for both the subjects analysis [$t(19) = -2.107, p < 0.05$] and the items analysis [$t(5) = 4.60, p < 0.001$].

Many studies on categorical perception have used a second measure as evidence for the presence of categorical perception. Typically, the discrimination data is correlated with predicted performance calculated on the results of the categorization task.³ The formula we used for deriving the predicted performance is adapted from that used by

³In summary, the function used to calculate the predicted discrimination performance for each face pair was the sum of the mean discrimination performance to the pairs of face images on either end of the gender continuum and 0.3 of the identification

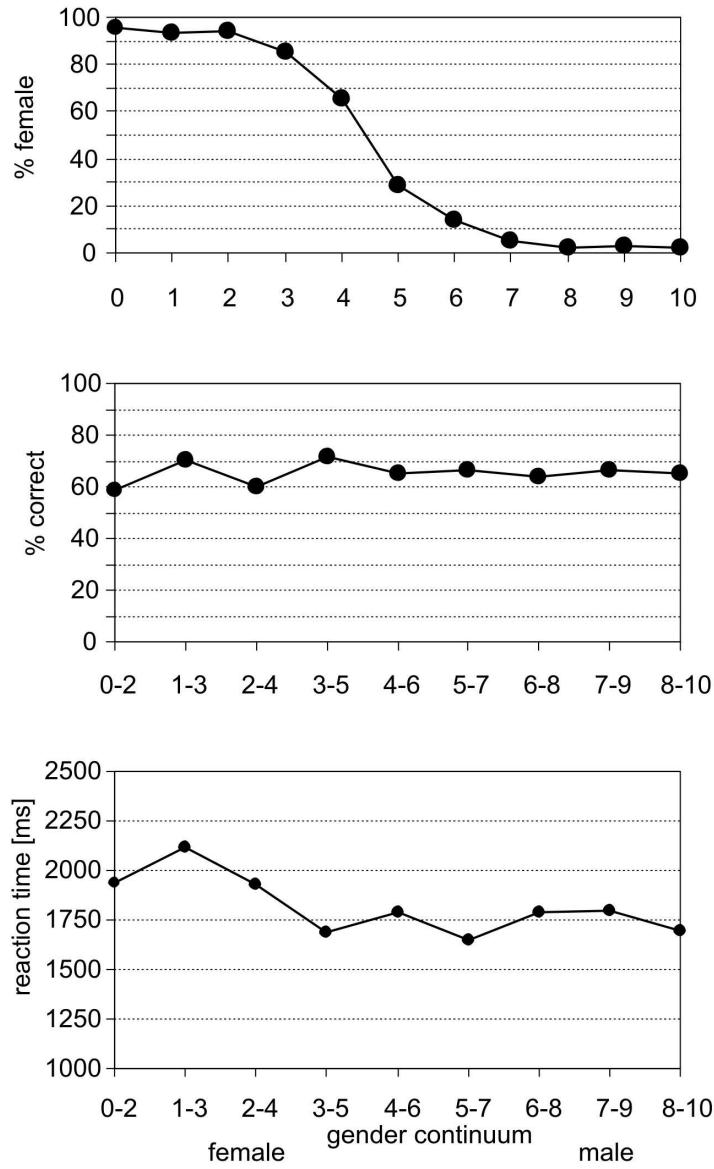


Figure 5: Plots showing data from Experiment 3: (upper row) mean categorization data, (middle row) mean discrimination data and (lower row) mean reaction times in the discrimination task.

Calder et al. [1996] and Liberman et al. [1957]. We conducted a Pearson correlation between the observed and predicted discrimination performance. The correlation was not significant, [$r = 0.3279$, n.s.].

Reaction times. The reaction time across all trials was 1821 ms. We conducted a one-way ANOVA on the mean reaction times to the different face image pairs in the discrimination task (see Figure 5 (lower row)). We found a significant effect of subjects [$F(8,152) = 2.203$, $p < 0.05$] and of items [$F(8, 40) = 2.705$, $p < 0.05$]. Post-hoc Newman Keuls analyses revealed that the reaction times to the image pairs 1-3 were significantly slower than reaction times to image pairs 5-7 and 8-10 ($p < 0.05$). Thus the variations in reaction times were not related to the CP effect found in the discrimination responses.

difference for each pair of images tested. See Newell and Bülthoff [2002], Calder et al. [1996] and Liberman et al. [1957] for further details.

5.3 Discussion

Categorization task. The shape of the categorization curves in Figure 6 indicates that categorization performance was improved for familiar faces at the female end of the continua; the endpoint female faces as well as morphs at the female end of the continua were more often categorized as female than in Experiment 1. The same endpoint female faces were judged to be female in 95.8 of the trials as compared to 76.8% in Experiment 1. At the male end of the continua, endpoint male faces were judged to be male in 97.8% of the trials as compared to 100% in Experiment 1. The subjective gender boundary was between faces 4 and 5 for familiar faces, which is closer to the physical image-based gender boundary (i.e. face 5) than for unfamiliar faces (the gender boundary was between faces 3 and 4 in Experiment 1). Thus the male bias was reduced, along with the better categorization for faces at the female end of the continua.

The improved categorization performance for female faces is explained by the preceding familiarization task; but the experimental design might also have contributed to improve performance. In this experiment, all morph images corresponding to a single face continuum were presented together in one of the six experimental blocks while all face combinations were mixed up in all blocks in Experiment 1. Thus, blocking the face stimuli may have promoted better effects of learning and consequently CP.

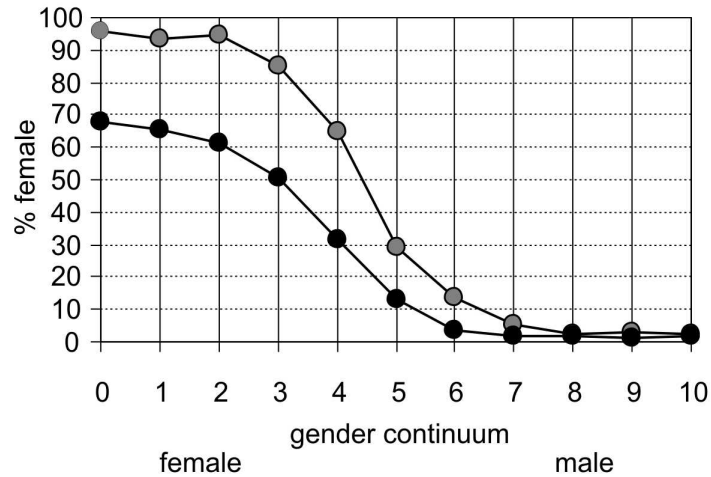


Figure 6: Plot showing the categorization curves of Experiment 1 and Experiment 3 superimposed for comparison. The grey line represents responses to the familiar faces and the black line represents responses to the unfamiliar faces.

Discrimination task. The results of the discrimination task show a definite categorical effect after familiarization that was not present in Experiment 1. Face pairs straddling the gender boundary were significantly better discriminated than end pairs of face images. Levin and Beale [2000] have shown that CP for the identity of newly learned faces exists after short term learning of the endpoint faces. In the first experiment of their study participants viewed pairs of faces and were asked which one was more similar to one of two faces they saw at the beginning of the testing block. Therefore, face combinations were blocked in their experiment. In our experiment we used a procedure somewhat similar to that of Levin and Beale, i.e. we tested each face combination separately in different blocks. Our procedure differs in that the observers were not familiarized with the endpoint faces prior to each experimental block but instead they were familiarized with all the faces grouped by gender before the whole experimental session.

We were concerned however whether the categorical effect measured in this experiment was a result of the familiarization with the identity or with the gender of the endpoint faces. Our face stimuli, like face stimuli used by other authors (Campanella et al. [2001]), confounds information relevant to gender with information related to identity. For example, by morphing a male face with a female face we also morphed the identity of the male face (e.g. Kurt) with the female identity (e.g. Heidi). Therefore, effects of CP may emerge due to idiosyncratic differences between the faces of e.g., Kurt and Heidi and not between male and female characteristics per se.

Furthermore, the blocked presentation might have favored the use of identity related information. The absence of a significant correlation between the observed and predicted discrimination performance might indicate that participants performed the discrimination task by using facial information pertaining to the identity of the faces, whereas the design of the categorization task obliged participants to perform this task using facial information pertaining to the gender of the faces only. Thus, we speculate that the lack of a significant correlation may be due to different information underlying different tasks. We controlled for possible effects of identity in Experiment 4.

Although an analysis of the discrimination data confirms the presence of a categorical effect, in our overall graph no peak is visible. Our training might not have been optimal for strong CP of gender or identity. Interestingly, recent results (Angeli et al. [2001]) suggest that categorical perception of identity might not be present when both endpoint faces are typical faces. Our face images were derived from typical faces but differed not only in identity but also in gender. Clearly the nature of the categorical effect measured in Experiment 3 remains an open question.

6 Experiment 4a and 4b

In the following experiment we investigated the effect of CP in male and female faces when identity information was controlled. Using the algorithm of Blanz and Vetter [1999] we created new gender continua in which the endpoint faces were of different gender but shared the same similar facial features. In Experiment 4a we created a continuum between a male and a female version of the average face. In Experiment 4b face stimuli created from continua between an original female face and a male face derived from this female face were used. With these face stimuli we could investigate whether face gender was perceived categorically when there was no change of characteristic facial features related to identity. In these experiments the stimuli were unfamiliar to the participants; the effect of familiarization will be investigated in Experiment 5.

6.1 Method

Participants. Fourteen students (4 males and 10 females) took part in Experiment 4a, and 12 participants (3 males and 9 females) took part in Experiment 4b.

Stimuli. For these experiments we used what we term "gender only" continua. The facial variations in our morph sequences were solely determined by the gender differences between the endpoint faces and not by any differences in identities. (Identity means here the unique characteristic features of a face). This was achieved in Experiment 4a by creating a gender continuum between a male and a female version of the average face and in Experiment 4b by creating continua between six original female faces and their corresponding male faces which were derived from these female faces.

In *Experiment 4a*, the neutral average face was calculated by linearly combining all 200 faces of the database. Differences between the 100 male and 100 female faces of the database were computed to generate a general "gender vector" that was free of individual facial variations and represented gender-related differences between all faces of both categories. In the context of a computationally derived face space the gender trajectory of the average face continuum passed through the average neutral face and joined the female average face on one side of the gender boundary to the average male face on the other side. Morphed average faces that vary in "gender strength," lay along this line (See Figure in Appendix for further explanation). The 11 stimuli of the continuum are represented in Figure 7 (upper row); they were all gender variations of the average face, thus all were synthetic faces computed from our face database. The middle image of the continuum represents the neutral average face.

For *Experiment 4b*, we generated what we call "one-identity" gender continua. Each of these continua was based on a single individual face. In a morphing procedure, all features of a female face were transformed (masculinized) using the gender vector described above. We computed the corresponding male faces of six female faces. Six one-identity face combinations; each based on a different female face, yielded a total of 66 stimuli. Figure 7 (6 lower rows) shows the six original female faces, their computed equivalent male faces and the morphed faces in-between the original female and derived male face.

Design and Procedure. *Experiment 4a* was based on a within subjects design. The discrimination task was based on a XAB match-to-sample paradigm similar to those used in the previous experiments. Pilot studies revealed



Figure 7: Illustration of the two types of gender-only continua used in Experiment 4: (upper row) *average-based* continuum used in Experiment 4a, (all other rows) the six *one-identity* continua used in Experiment 4b. In the upper row, the average female face is the first face on the left; the average male face is the last face on the right. In all other rows, the six original female faces are on the left. Gender continua in the second and third upper row were used in Experiment 5. For more explanation, see text.

that participants' performance was around chance level when the AB images were 2 steps apart. Therefore in this experiment the AB images were 3 steps apart. There were eight AB stimuli pairs (i.e. $0 - 3, 1 - 4 \dots 7 - 10$). Each image pair was presented eight times, yielding 64 trials (one gender continuum, 8 image pairs, repeated 8 times) presented in a one-block session. The procedure was the same as in previous experiments except that the X stimulus was shown for 1000 ms instead of 750 ms. In the following categorization task all stimuli were presented five times, giving a total of 55 trials in one session (one gender continuum, 11 face images, repeated 5 times). In all other ways the design and procedure were the same as in Experiment 1. Participants needed 5 to 9 minutes to complete this experiment.

Experiment 4b was also based on a within subjects design. Pilot studies had revealed that participants were around chance level when the AB images were 3 steps apart in the XAB match-to-sample paradigm. Therefore we chose an easier discrimination task. Here participants performed a simultaneous same/different task (see Calder et al. [1996] and the two face images in each trial were 3 steps apart. There were eight "different" stimuli pairs (0 – 3, 1 – 4 ··· 7 – 10) and eleven "same" pairs (0 – 0, 1 – 1 ··· 10 – 10). Trials were blocked by face combination. There were 65 trials per face combination with 33 same trials and 32 different trials. Thus there were 390 trials per session (6 gender continua, 65 trials per face combination). A fixation cross preceded the pair of face stimuli for 250ms in each trial. The pair of stimuli were presented simultaneously about 3 cm left and right of the fixation point and remained on the screen until the participant pressed a response button of the button box. An inter-trial interval of 500ms followed the participant's response. Participants had to indicate whether both images were the same or different by pressing the left or right response button respectively. There was always a short practice block of 8 trials before starting the test blocks. In all other aspects the design and procedure were the same as in the other experiments. For the following categorization task each face image was presented four times, giving a total of (6 continua x 11 images repeated 4 times) 264 trials in one session. The stimuli were blocked by continua. In all other ways the design and procedure were the same as in Experiment 1. Participants needed about 55 minutes to complete this experiment.

6.2 Results

6.2.1 Experiment 4a

The mean number of correct responses made to the XAB task and the mean percentage of female decisions to the gender categorization task are shown in Figures 8 (upper and middle row).

Categorization task. In this experiment, the subjective category boundary lay between face images 4-5.

Discrimination task. Overall performance over all face pairs was 61.5%. We conducted a paired t-test using category position as the factor. As before, the category position factor refers to the discrimination performance at pairs of face images that lay at either end of the gender continua (i.e. the average performance to face image pairs 0-3 and 7-10) compared to image pairs that straddle the category boundary (i.e. faces images 3- 6). The mean number of correct responses for the end pairs of face images was 67.2% and for the images straddling the category boundary it was 67.8%. We found no significant effect of category position [$t(13) = -0.1375$, n.s.]. (Items analyses were not conducted as there was only one gender continuum tested.).

Reaction times. The mean reaction time across all trials was 2399 ms. The average reaction times for each face pair are plotted in Figure 8 (lower row). We conducted a one-way ANOVA on the mean reaction times to the different face pairs in the discrimination task. We found no significant effect of face pair [$F(7,91) = 1.94$, n.s.].

6.2.2 Experiment 4b

The mean percentage of female decisions to the gender categorization task is shown in Figure 9 (upper row). The discrimination scores on the same/different task were converted to d' scores (a sensitivity measure). Figure 9 (middle row) shows the mean d' scores across all participants and all continua.

Categorization task. As before, the point of subjective category boundary was determined by the results of the categorization task. In this experiment, the subjective category boundary lay between face images 3 and 4.

Discrimination task. Overall performance was 71.1% over all different pairs and 64.5% over all same pairs. We conducted a paired t-test on the mean d' scores to the face image pairs at the end of the continuum (face images 0-3 and 7-10) and the face images that straddled the category boundary (images 3-6). We found no effect of category position on the d' data [$t(11) = 0.333$, n.s.].

Reaction times. Participants reported that they had great difficulties in performing the task. This is reflected in their reaction times. The average reaction times for all face pairs are plotted in Figure 9 (lower row). The mean reaction time to the same trials was 7.8 seconds and to the different trials was 7.1 seconds. We conducted separate one-way ANOVAs on the mean reaction times to the same image pairs and the different face image pairs in the

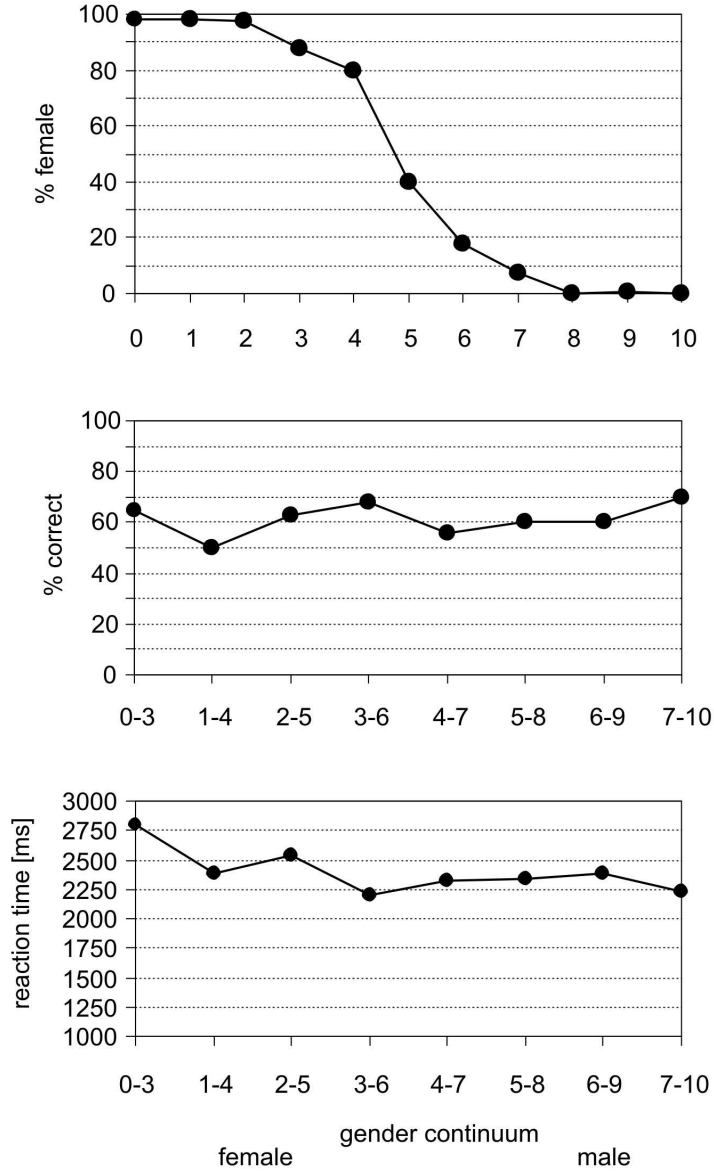


Figure 8: Plots showing data from Experiment 4a: (upper row) mean categorization data, (middle row) mean discrimination data and (lower row) mean reaction times in the discrimination task.

discrimination task. As expected in regard to the very long reaction times, we found no significant effect of reaction times for the same trials [$F(10,110) = 0.83$, n.s.] or for the different trials [$F(7,77) = 1.54$, n.s.].

6.3 Discussion

Categorization task. In *Experiment 4a* none of the face images corresponded to the face of a real person. The neutral average face was computed as a linear summation of all faces in our database, while the computed differences between male and female features of the faces of the database were used to generate the male and female average faces and to calculate in-between or morphed faces. The results of the categorization task indicates that gender information is present in the synthetic stimuli and that participants could extract this information to perform the categorization task as well as when the continua endpoints were formed by faces of real persons. These results are an important demonstration of the validity of the algorithm that we used throughout this study and of the quality of the database that was used to produce gender variation of the average face. Furthermore the

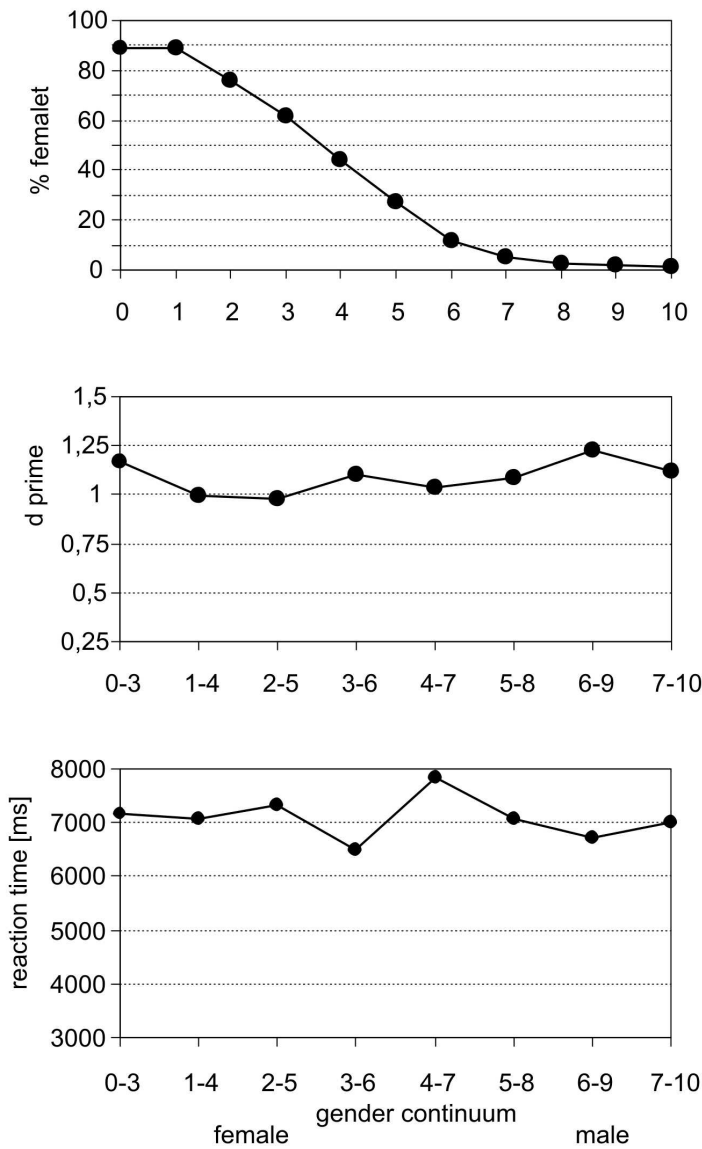


Figure 9: Plots showing data from Experiment 4b: (upper row) mean categorization data, (middle row) mean d' discrimination data and (lower row) mean reaction times in the discrimination task.

male bias observed in our previous experiments is almost absent here. By virtue of the averaging process, all faces of the average gender continuum have smoother features than the morphs created from two real individuals. The averaging process might have rendered the faces more feminine to the observers (Simons [1997]). It was recently noted that because morphs generally have smoother features than original faces, this might influence the underlying psychological representation of morphs (Busey [1998]).

In *Experiment 4b* only the faces at the female endpoint of each continuum corresponded to the features of a real person, and all other faces were artificially created. The shape of the response curve does not form a categorization step as clearly as in Experiment 4a. Nevertheless the curve suggests that participants could identify the proportion of both genders in the images and could classify the faces by their gender mixture accordingly. Here a male bias in responses is again present to the same degree as was found in Experiment 1.

Discrimination task In *Experiment 4a*, the lack of characteristic features related to identity made the average face stimuli difficult to distinguish from each other. Although the faces in each test trials were 3 steps apart, discrimination performance was not better than in the previous experiments. An analysis of the individual results indicated that discrimination performance to the face pair that straddles the category boundary was not better than to face pairs at the endpoints of the continua.⁴ The analysis of the reaction times likewise showed an absence of a categorical effect.

In *Experiment 4b*, again, the face pairs straddling the gender boundary were not significantly easier to discriminate than within-category pairs. Because of the difficulty of the task observed in pilot studies and in order to ensure that participants could perform the task, we did not set a time limit. As a consequence, the reaction times were extremely slow.

In *Experiments 4a and 4b*, we were surprised to realize how difficult it was for the participants to discriminate between faces from the "gender-only" continua. Although we are expert at deciphering slight changes in facial expression, we are obviously not trained to discriminate between faces differing in gender quality only. The fact that participants had more difficulties with one-identity continua than with the average gender continuum was not expected either. More research would be needed to determine whether this difference in discrimination difficulty between one-identity continua and the average continuum is relevant and whether this difference is related to the presence or absence of characteristic "identity" features.

Our findings suggest that gender-related changes in facial features are less obvious to the observers than identity-related changes.⁵ This is especially clear when we compare the performance of participants with the one-identity continua used in Experiment 4b to the results obtained by Beale and Keil [1995] with face continua in which the gender remained the same but the identity varied. Participants in the first experiment of the Beale and Keil study could perform a XAB match-to-sample task with AB pairs that were 2 steps apart while participants in Experiment 4b were at chance even when the one-identity face pairs were 3 steps apart. The results of Beale and Keil that we refer to were obtained for well-known faces but unfortunately they used a different paradigm when they tested unfamiliar faces thus it is difficult to directly compare our studies. Differences or interactions between the perception of identity and gender have been investigated by few authors (with the exception of Goshen-Gottstein and Ganel [2000]; Baudoin and Tiberghien [2002] and needs further empirical, as well as theoretical scrutiny.

What is important for our investigation here is that again participants did not perform the discrimination task in a categorical way. This result confirms the findings in Experiment 1, i.e. a lack of CP for gender, and demonstrates that it was not facial information related to the change of identity that hindered CP for gender in Experiment 1. The results we obtained in Experiment 4 suggest again that gender generally is not perceived categorically in faces.

Due to the difficulty of the discrimination task for the face stimuli used here and the general difficulty to obtain categorical perception of gender throughout this study, we wondered whether pure gender information in faces was at all available for categorical perception. We addressed this question in the next experiment.

7 Experiment 5

In Experiment 4, participants were not familiar with the face stimuli. In the present experiment we treated gender categories as unfamiliar categories that participants had to learn. Consequently, we trained participants to categorize the face stimuli by gender (see Goldstone [1994]; Goldstone et al. [2001] for similar procedures) prior to testing and investigated whether training could induce categorical perception of gender. Many studies have shown that CP can be induced for unfamiliar categories after proper training (see Adamson and Sowden [2000], Adamson and Sowden [2001]; Goldstone [1994]; Goldstone et al. [2001]).

⁴We chose for the analysis performance to the "best" of three possible pairs of faces that straddle the boundary, i.e. the pair for which participants showed on average the best discrimination performance. Because of the bigger difference between the faces of a pair (3 steps instead of 2 steps), other face pairs straddle the category boundary.

⁵As an interesting aside, we used the same method to masculinize the head of one of the authors (F. N.). The resulting face was clearly identified by the other family members as the face of one of her.

7.1 Method

Participants. Eighteen participants took part in the following experiment. Thirteen of the participants were female. One female and two male participants could not meet the learning criterion of 90% correct in the learning phase and were excluded from further analyses.

Stimuli. In this experiment we only used two of the six face continua used in Experiment 4b. We wanted to get enough measurements on a few face pairs whilst not over-burdening our participants with this difficult task.

Design. Participants went through a training session before performing the discrimination and categorization tasks similar to those described in Experiment 4b. The training phase consisted of the same forced-choice gender categorization task as in the test, except that participants were given appropriate visual and auditory feedback on the accuracy of their response on each trial. A response was labeled as correct every time a participant categorized any of the five images on the female (or male) side of the continuum as female (or male). Ten faces of each continuum were shown 10 times in the training phase; faces 0 to 4 (6 to 10) had to be labeled "female" ("male") by the participant for a correct response. Note that face image 5 was not presented because it lay on the gender boundary. Thus there were 100 trials per training session. The order of the trials within each session was randomized across participants. There were two training blocks, one for each one-identity continuum. The order of presentation of the two training blocks was randomized. Participants conducted an experimental session after each training block. In the test phase, participants were presented with the same stimuli as in the training phase. The design of the discrimination and the categorization tasks was the same as in Experiment 4b except that image 5 was not used in the categorization task. There were 65 trials in each discrimination task and 30 trials in each subsequent categorization task. Participants received a self-timed break between each block.

Procedure

Training session. Each training trial consisted of the following sequence of events: a fixation cross was shown for 250 ms, followed by a training face image that remained on the screen and to which the participant responded male or female. As soon as the participant responded, feedback in the form of a text display stating the correctness of the participants' decision was given. This feedback was displayed for 750 ms. Visual feedback was coupled with auditory feed-back. Depending on the correctness of the response, one of two different beeps was presented. An inter-trial interval of 500ms followed the feedback display.

Experimental session. The procedure of the experimental session followed that of Experiment 4b. Participants needed around 40 minutes to perform the experiment.

7.2 Results

Training phase. The average performance of all participants, calculated over both training periods, was 92.2% correctly classified trials.

Test phase. The mean discrimination performance for the same/different task and the mean percentage of female decisions to the gender categorization task are shown in Figures 10 (upper and middle row).

Categorization task. The perceived gender boundary was very close to face image 5, thus close to the given gender boundary. Participants classified 94% of all faces belonging to the female category (face 0 to face 4) as female and 94% of all male faces (face 6 to face 10) as male. Almost all categorization errors occurred to the two faces closest to the gender boundary in each gender category. For comparison, participants without training classified correctly 81% of the female faces and 91% of the male faces of the same continua in Experiment 4b.

Discrimination task. Overall performance was 58% correct to the different-face pairs and 66% correct to the same-face pairs. The discrimination scores on the same/different task were converted to d' scores. We conducted a paired t-test on the mean d' scores to the face image pairs at the end of the continuum (face images 0-3 and 7-10) and the face images that straddled the category boundary (images 4-7). We found an effect of category position on the d' data [$t(15) = -2.433, p < 0.05$] in the subjects analysis.

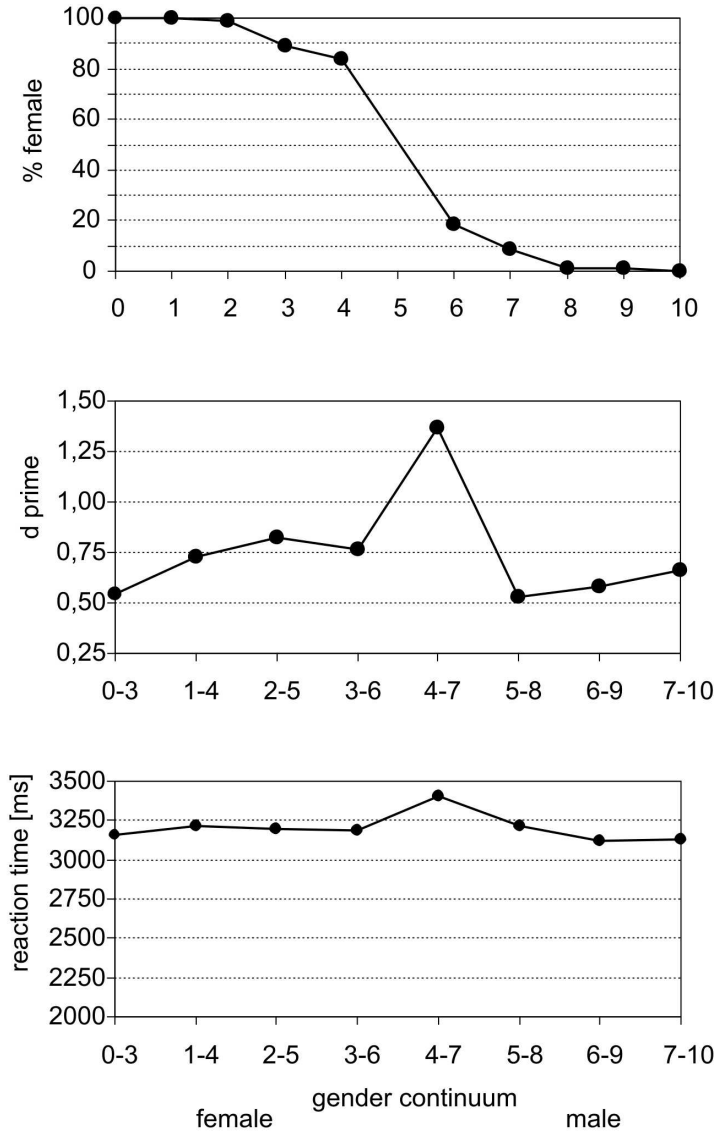


Figure 10: Plots showing data from Experiment 5: (upper row) mean categorization data, (middle row) mean d' discrimination data and (lower row) mean reaction times in the discrimination task.

We used the same method as in Experiment 3 to derive predicted performance on the results of the categorization task. Note that we have no categorization data for face image 5. We conducted a Pearson correlation between the observed and predicted discrimination performance.

The correlation was close to being significant, [$r = 0.7459$, $p = 0.054$].

Reaction times. Participants did not complain about the difficulty of the task in this experiment. This is reflected in their reaction times, i.e., here participants needed about half the time to respond than participants in Experiment 4b. The average reaction times for all face pairs are plotted in Figure 10 (lower row). The mean reaction time to the same-face trials was 3572 ms and to the different-face trials was 3203 ms. We conducted separate one-way ANOVAs on the mean reaction times to the same image pairs and the different face image pairs in the discrimination task. We found no significant effect of reaction times for the same trials [$F(10,130) = 1.52$, ns] or for the different trials [$F(7, 98) = 0.27$, ns]. Likewise, a t-test between the mean RT to the face image pairs at the end of

the continuum and the face images that straddled the category boundary showed no significant differences [$t(14) = -1.388$, n.s. for same pairs and $t(14) = -1.077$, n.s. for different pairs].

7.3 Discussion

Categorization task. In comparison to the results of Experiment 4b where the participants were not familiarized with the stimuli prior testing, classification of faces by gender was improved with training and this categorization improvement was still present after participants performed the discrimination task.

Discrimination task. Participants improved their sensitivity to facial information pertaining to gender after category training but only for the face pairs straddling the category boundary. The accuracy for the face pairs straddling the gender category was increased (d' was 1.03 before training and 1.37 after training). The response times were drastically reduced for all face pairs compared to those found in Experiment 4b. Response times of the trained participants were twice as fast compared to the response times of participants without training. Nevertheless participants did not answer any faster when they had to discriminate between faces that straddled the category boundary than when both faces belonged to the same category. Thus, whilst on the one hand trained participants were much faster they were less accurate for within-category pairs and more accurate for category-straddling pairs compared to untrained participants. We also found that the correlation between predicted and actual performance was close to significant. We think the only reason it is not significant is because of the missing data point in the predicted function. At least, the correlation between predicted and actual performance is much better than in Experiment 3. We did not investigate further the difference between the results of Experiment 4b and 5 as it was not our purpose to examine whether category training induced CP by compressing within-category distances and/or by expanding between-category distances (see Harnad [1987] amongst others for more details on this point). The important result was that CP for gender was present after gender category training on face stimuli which differed on gender characteristics only.

8 General Discussion

In Experiments 1, 2 and 4 we found that unfamiliar male and female faces were not perceived as discrete categories. In these experiments there was no evidence for a categorical effect in gender discrimination. In contrast, when participants were trained on the face stimuli then CP effects emerged (Experiments 3 and 5).

In Experiment 1 we tested for effects of categorical perception in unfamiliar faces, under presentation conditions which minimized learning of the endpoint stimuli. We found no evidence for CP. In Experiment 2 we repeated Experiment 1 but tested the effects of face texture and face shape separately on gender perception. We found that textural information was easier to use than shape for gender categorization. No evidence of categorical perception was present in the discrimination results for either shape changes only or texture changes only.

In Experiments 1, 2 and 3, identity and gender changed in the face stimuli, therefore the possibility remained that participants were extracting identity information in order to perform the tasks. In Experiment 4 we created a set of stimuli where only the gender information changed along each continuum. We used continua with gender changes of the average face and gender changes of individual faces as stimuli. Again we found no evidence of CP for unfamiliar faces.

In Experiments 3 and 5 we tested effects of learning on CP of gender. First, we familiarized participants with the endpoint (i.e. original) faces prior to testing in Experiment 3. We found that the discrimination performance displayed the characteristics of categorical perception. However, although discrimination performance was better to face pairs from across rather than within a category, these effects were quite weak because the predicted discrimination performance derived from the categorization data did not correlate with the observed discrimination data. In Experiment 5, we trained participants to classify the faces of gender-only continua using feedback. Categorical perception for gender was found after training. Furthermore, the correlation between the predicted and actual discrimination performance was almost significant.

Contrary to our findings, Campanella and colleagues reported effects of CP for the gender of unfamiliar faces in a recent study (Campanella et al. [2001]). Differences between the morphing procedure and experimental designs may account for the differences across the two studies. Campanella et al. used face combinations between female

and male faces, resulting in continua containing both gender and identity changes. Our results question their claim that identity does not play any role in their finding a categorical effect.

In our study we found that the gender of unfamiliar faces is not perceived categorically. With training, however, CP effects emerge. Training has been shown to affect CP in many types of stimuli e.g. for less complex stimuli as reported by Goldstone [1994], Goldstone et al. [1996], and for more complex stimuli such as histological slides (Adamson and Sowden [2000], Adamson and Sowden [2001] and faces (Goldstone et al. [2001])). Therefore, we conclude, that face gender is not a naturally occurring perceptual category. Maybe this is related to the fact that we usually do not deal with faces in isolation when making gender judgements; even in portraits other cues to gender like clothing, accessories, make-up, hairstyle and hairline are present. Thus we may not have learned to use facial information categorically from the real world. Although we can tell male from female faces, we do not perceive face gender categorically without training.

The absence of a categorical effect is surprising. Other important information displayed by faces (e.g. expressions and identity) has been proved repeatedly to be perceived categorically (Beale and Keil [1995]; Levin and Beale [2000]; Young et al. [1997]). Clearly, despite its enormous importance for social interactions we have not learned to deal with the gender of faces very effectively. We speculate, therefore, that this might be the reason why many women enhance the gender characteristics of their faces by, for example, wearing make-up and plucking their eyebrows (Bruce et al. [1993]).

Acknowledgments

We are grateful to Volker Blanz and Thomas Vetter for their help with the face stimuli and the morphing technique and to Ian M. Thornton and Heinrich H. Bülhoff for insightful comments. The second author was supported by Max Planck Society and Trinity College Dublin Research Fund.

Appendix

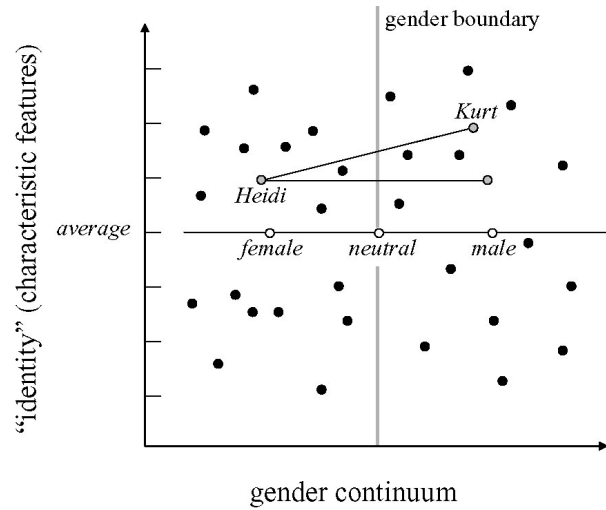


Figure 11: Appendix

An illustration of face space and manipulations of face gender. If each 3D face of the database is described by a vector containing the geometrical and textural data (see Blanz and Vetter [1999] for details), each face will have a location vector in a multidimensional space called face space (Valentine [1991]). Each dimension of this face space represents a different facial characteristic. Within this continuous space, it is possible to traverse a path from any face to any other face, morphing through locally similar faces along that path. Here we show a very simplified 2D representation of that face space with faces segregated by gender on each side of the gender boundary. The average of all faces is the neutral or average face which is at the center of the face space subtended by the faces of the database. The average male face and average female face are placed correspondingly. The mean gender differences between male and female faces are represented by the *gender vector* joining the average female face to the average male face. By manipulating the position of a face along, what we term, its gender trajectory given by the gender vector, one can systematically vary this face's gender alone. Manipulations of the average face along the gender vector were done in Experiment 4a. The gender of any individual face (e.g. Heidi) can be manipulated along the face's gender trajectory to the corresponding masculinized face. Note that the characteristic features (identity) of this face remains constant along this vector. Individual faces were manipulated along the gender vector in Experiment 4b and Experiment 5. By contrast a *gender and identity trajectory* is shown between two faces of different identities (here Heidi and Kurt). Manipulations of individual faces along a gender and identity vector were done in Experiment 1 and Experiment 3.

References

- H. Abdi, D. Valentin, B. Edelman, and A. J. O'Toole. More about the difference between men and women: Evidence from linear neural network and principal component approach. *Perception*, 24:539–562, 1995.
- S. E. Adamson and P. T. Sowden. Induced categorical perception of novel, real-world, complex stimuli. *Perception*, 29, Supplement, 91c, 2000.
- S. E. Adamson and P. T. Sowden. Does induced categorical perception of novel, real-world, complex stimuli result from perceptual learning? *Perception*, 30, Supplement, 102c, 2001.
- A. Angeli, J. Davidoff, and T. Valentine. Distinctiveness induces categorical perception of unfamiliar faces. *Perception*, 30, 58, 2001.
- J. Y. Baudoin and G. Tiberghien. Gender is a dimension of face recognition. *Journal of Experimental Psychology-Learning, Memory, and Cognition*, 28:362–365, 2002.
- J. M. Beale and F. C. Keil. Categorical effects in the perception of faces. *Cognition*, 57:217–239, 1995.

- V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.
- V. Blanz. Automatische Rekonstruktion der dreidimensionalen Form von Gesichtern aus einem Einzelbild, 2000.
- M. H. Bornstein and N. O. Korda. Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research*, 46:207–222, 1984.
- M. H. Bornstein. Perceptual categories in vision and audition. In Stevan Harnad, editor, *Categorical perception: The groundwork of cognition*, pages 287–300. Cambridge University Press, New York, NY, USA, 1987.
- R. M. Boynton. *Human Color Vision*. Holt, Rinehart & Winston, New York, 1979.
- V. Bruce, A. M. Burton, E. Hanna, P. Healley, O. Mason, A. Coombe, R. Fright, and A. Linney. Gender determination: How do we tell the difference between male and female faces? *Perception*, 22:131–152, 1993.
- V. Bruce and S. Langton. The role of pigmentation and shading information in recognising the gender and identities of faces. *Perception*, 23:803–822, 1994.
- V. Bruce. Influence of familiarity on the processing of faces. *Perception*, 15:387–397, 1986.
- E. M. Burns and W. D. Ward. Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America*, 63:456–68, 1978.
- M. Burton, V. Bruce, and N. Dench. What’s the difference between men and women? Evidence from facial measurement. *Perception*, 22:153–176, 1993.
- T. Busey. Physical and psychological representations of faces: Evidence from morphing. *Psychological Science*, 9:476–482, 1998.
- A. J. Calder, A. W. Young, D. I. Perrett, N. L. Etcoff, and D. A. Rowland. Categorical perception of morphed facial expressions. *Visual Cognition*, 3:81–117, 1996.
- S. Campanella, A. Chrysochoos, and R. Bruyer. Categorical perception of facial gender information: Behavioural evidence and the face-space metaphor. *Visual Cognition*, 8:237–262, 2001.
- Y. Cheng, A. J. O’Toole, and H. Abdi. Classifying adults’ and children’s faces by sex: Computational investigations of subcategorical feature encoding. *Cognitive Science*, 25:819–838, 2001.
- C. D. Clemente. *Gray’s anatomy of the human body. 30th Edition*. Lea & Febiger, Philadelphia, 1985.
- J. D. Cohen, B. MacWhinney, M. Flatt, and J. Provost. Psyscope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments & Computers*, 25:257–271, 1993.
- B. de Gelder, J. Teunisse, and P. J. Benson. Categorical perception of facial expressions: Categories and their internal structure. *Cognition & Emotion*, 11(1):1–23, 1997.
- P. D. Eimas, J. L. Miller, and P. W. Jusczyk. On infant speech perception and the acquisition of language. In S. R. Harnad, editor, *Categorical perception: The groundwork of cognition*, pages 161–195. Cambridge University Press, Cambridge, U.K.
- P. Ekman. Strong evidence for universals in facial expressions: A reply to russell’s mistaken critique. *Psychological Bulletin*, 115:268–287, 1994.
- N. L. Etcoff and J. J. Magee. Categorical perception of facial expressions. *Cognition*, 44:227–240, 1992.
- R. L. Goldstone, Y. Lipka, and R. M. Shiffrin. Altering object representations through category learning. *Cognition*, 78:27–43, 2001.
- R. L. Goldstone, M. Steyvers, and K. Larimer. Categorical perception of novel dimensions. In *Proceedings of the eighteenth annual conference of the Cognitive Science Society*, pages 243–248, La Jolla, CA, 1996. Lawrence Erlbaum Associates.
- R. L. Goldstone. Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123:178–200, 1994.

- Y. Goshen-Gottstein and T. Ganel. Repetition priming for familiar and unfamiliar faces in a sex-judgment task: Evidence for a common route for the processing of sex and identity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26:1198–1214, 2000.
- S. R. Harnad. *Categorical Perception: The groundwork of cognition*. Cambridge University Press, Cambridge, U.K., 1987.
- H. Hill, V. Bruce, and S. Akamatsu. Perceiving the gender and race of faces: The role of shape and colour. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 261:367–373, 1995.
- M. Intons-Peterson. *Children's concepts of gender*. Ablex, Norwood, NJ, 1988.
- D. T. Levin and J. M. Beale. Categorical perception occurs in newly learned faces, cross-race faces, and inverted faces. *Perception and Psychophysics*, 62:386–401, 2000.
- A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological Review*, 74:431–61, 1967.
- A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54:358–368, 1957.
- F. N. Newell and H. H. Bülthoff. Categorical perception of familiar objects. *Cognition*, 85:113–143, 2002.
- A. J. O'Toole, S. Edelman, and H. H. Bülthoff. Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research*, 38:2351–2363, 1998.
- A. J. O'Toole, T. Vetter, N. F. Troje, and H. H. Bülthoff. Gender classification is better with three-dimensional head structure than with image intensity information. *Perception*, 26:75–84, 1997.
- E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.
- F. Simons. Sexual discrimination on prototype face, 1997. Poster, Réunion de la Société Belge de Psychologie, Bruxelles.
- N. F. Troje and H. H. Bülthoff. How is bilateral symmetry of human faces used for recognition of novel views? *Vision Research*, 38:79–89, 1996.
- T. Valentine. A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology*, 43A:161–204, 1991.
- R. L. De Valois and K. K. De Valois. Neural coding of color. In *Handbook of perception*, volume 5, pages 117–166. Academic Press, New York.
- T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):733–742, 1997.
- H. H. Wild, S. E. Barrett, M. J. Spence, A. J. O'Toole, Y. D. Cheng, and J. Brooke. Recognition and sex categorization of adults' and children's' faces in the absence of sex stereotyped cues. *Journal of Experimental Child Psychology*, 77:261–299, 2000.
- A. W. Young, D. A. Rowland, A. J. Calder, N. L. Etcoff, A. Seth, and D. I. Perrett. Facial expression megamix: Test of dimensionality and category accounts of emotion recognition. *Cognition*, 63:271–313, 1997.
- L. A. Zebrowitz. *Reading Faces: Window to the Soul?* Westview Press, Boulder, CO, 1997.