Can low level image differences account for the ability of human observers to discriminate facial identity?

Danelle A. Wilbraham Ohio State University, USA

James C. Christensen Ohio State University, USA

Aleix M. Martinez Ohio State University, USA

James T. Todd Ohio State University, USA

A fundamental difficulty for image- or appearance-based models of face recognition is to distinguish variations in image structure between two different individuals from those that can occur for a given individual due to changes in lighting, facial expression, or pose. The research described in the present article was designed to examine how human observers are able to cope with this problem. In two experiments, observers performed either a match-to-sample task (Experiment 1) or same–different identity judgments (Experiment 2) for photographs of unfamiliar individuals. A key aspect of these studies is that the matching or same stimulus pairs were never identical; that is to say, they always differed in terms of facial expression or the pattern of illumination. In order to provide a quantitative assessment of appearance-based models, we also measured the optical differences for each pair of same or different images using a variety of possible distance metrics based on the pattern of pixel intensities or wavelet decompositions. These difference measures were then correlated with the accuracy of observers' judgments for each individual stimulus pair. The results clearly show that human observers can readily distinguish relevant from irrelevant image changes in comparisons of facial identity, but that this performance cannot be explained by any of the appearance-based models we tested.

Keywords: face recognition, computational modeling, 3D surface and shape perception

Citation: Wilbraham, D. A., Christensen, J. C., Martinez, A. M., & Todd, J. T. (2008). Can low level image differences account for the ability of human observers to discriminate facial identity? *Journal of Vision, 8*(15):5, 1–12, http://journalofvision.org/8/15/5/, doi:10.1167/8.15.5.

Introduction

The ability of human observers to reliably identify faces is a truly remarkable phenomenon. Despite the fact that all human faces have a similar overall structure, we are able to identify people from different vantage points, and with different patterns of illumination, facial expressions, hair styles, makeup, or clothing accessories, such as hats or glasses. We can also identify people after they undergo a growth spurt, gain or lose weight, or suffer the effects of aging. These observations suggest that the identity of an individual's face must be based on some remarkably abstract property that is somehow unaffected by all of the transformations that faces typically undergo in the natural environment.

There are two general approaches that have been described in the literature for how faces might be perceptually encoded within the human visual system.[1] One common hypothesis that we will refer to generically as the feature-based approach is that faces are represented by the local shapes of their distinctive features (e.g., the eyes, nose, mouth, and chin) and the spatial relationships among those features (Barton, Zhao, & Keenan, 2003; Cooper & Wojan, 2000; Sadr, Jarudi, & Sinha, 2003). The primary evidence to support this hypothesis is that face recognition is significantly impaired when images are edited to remove facial features or spatially rearrange them (Bruce & Young, 1998; Ellis, Shepherd, & Davies, 1979; Nachson, Moscovitch, & Umilta, 1995; Sinha, 2002a; Sinha & Poggio, 1996, 2002; Young, Hay, McWeeny, Flude, & Ellis, 1985). A feature-based approach is also the strategy that was first employed in the earliest computational models of face recognition within the field of machine vision (Craw, Ellis, & Lishman, 1987; Goldstein, Harmon, & Lesk, 1971; Kanade, 1973; Kaya & Kobayashi, 1972). Although these models are able to achieve satisfactory performance when facial features are extracted manually, their success has been limited by the inherent difficulty of developing robust algorithms for the automatic extraction of facial features under general viewing conditions (e.g., Brunelli & Poggio, 1993).

In an effort to circumvent this difficulty, other researchers have proposed an image- or appearance-based approach to face recognition that bypasses the problem of feature extraction altogether (Biederman & Kaloscai, 1997; Meytlis & Sirovich, 2007; Sirovich & Kirby, 1987; Tarr & Gauthier, 1998; Turk & Pentland, 1991). The basic idea of this approach is to represent images of faces as

doi: 10.1167/8.15.5 Received March 14, 2008; published November 17, 2008 ISSN 1534-7362 © ARVO

arrays of pixel intensities or wavelet outputs that are analogous to the response patterns of photoreceptors on the retina or simple cells in V1. Recognition is achieved by comparing images to stored templates using a suitable metric such as Euclidean distance. Because appearance-based representations generally result in an excessive number of dimensions, it is common for these models to employ a data reduction algorithm such as principal components analysis (PCA), which can reduce the number of dimensions by two or three orders of magnitude, yet still account for almost all of the variance among the set of images to be represented. The primary advantage of appearance-based models relative to feature-based approaches is that they are mathematically well specified and can therefore be implemented as actual working models without requiring human intervention for the extraction of meaningful features.

There is some research to suggest that these appearance-based algorithms might also be considered as viable models of human face recognition. One of the primary limitations of appearance-based algorithms is that they have difficulty coping with image differences that are irrelevant to an individual's identity, such as those resulting from changes in illumination, facial expression, or pose. Empirical studies have shown, however, that human facial identity judgments are also impaired by these irrelevant image changes (Braje, 2003; Braje, Kersten, Tarr, & Troje, 1998; Hill & Bruce, 1991, 1996; Hill, Schyns, & Akamatsu, 1997; Liu & Chaudhuri, 2002; O'Toole, Edelman, & Bülthoff, 1998; Tarr, Kersten, & Bülthoff, 1998; Troje & Bülthoff, 1998), thus suggesting that the performance of these algorithms is similar to that of human observers. Of particular interest in this regard is that line drawings of famous faces, which isolate the information that is most relevant for feature-based approaches, produce much lower recognition rates than is typically obtained with photographs (Benson & Perrett, 1994; Davies, Ellis, & Shepherd, 1978; Rhodes, Brennan, & Carey, 1987). Although these findings may appear at first blush to provide strong empirical support for an appearance-based model of human face recognition, the impact of this evidence is muddled by the absence of quantitative measures to evaluate differences among the facial images observers are asked to judge. The results show clearly that recognition is impaired by irrelevant image changes, but it has not yet been determined if the magnitude of these impairments is consistent with those that would be expected based on current computational algorithms.

There are two important issues that need to be considered in order to provide a quantitative evaluation of appearance-based algorithms as potential models of human face recognition. First, it is important to keep in mind that there are many possible methods for measuring image differences that have been described in the literature, and there have been no systematic studies to evaluate the extent to which they are consistent with one another. Thus, in order to provide a general assessment of appearance-based approaches, it is necessary to examine a reasonably broad sample of possible similarity metrics.

A second important issue for evaluating the psychological validity of face recognition models is to select a method for comparing their quantitative predictions with the performance of human observers. The most common procedure for accomplishing this goal in prior studies has been to compare the overall percentage of correct responses (Valentin, Abdi, Edelman, & O'Toole, 1997; Wallraven, Schwaninger, & Bülthoff, 2005). Note, however, that this is a relatively crude criterion, because it is possible for two models to achieve the same overall accuracy with quite different patterns of errors. A more stringent analysis for evaluating face recognition models is to compare their performance with human observers for all of the individual stimuli employed in an experiment in order to demonstrate if the relative difficulty among different stimulus items is the same for the model as it is for observers (Burton, Miller, Bruce, Hancock, & Henderson, 2001).

In light of these observations, the research described in the present article was designed to provide a quantitative assessment of the extent to which appearance-based models can account for the ability of human observers to distinguish images of different individuals under varying conditions of illumination, facial expression, or partial occlusion. A key aspect of these studies is that images depicting the same individual were never identical: They always differed in terms of facial expression or the pattern of illumination. To facilitate subsequent analyses, we also measured the overall similarity of each pair of images the observers were asked to judge using a wide variety of possible distance metrics based on the pattern of pixel intensities or wavelet decompositions. These difference measures were then correlated with the accuracy of observers' judgments for each individual stimulus pair.

Experiment 1

Methods

Apparatus

The experiment was controlled by a Dell Dimension 8300 computer with a 21-inch CRT display. The spatial resolution of the display was 640×480 pixels. This display subtended 32 by 24 degrees of visual angle when viewed from a distance of 76 cm. The timing of the experimental displays and response collection were controlled with E-Prime by Psychological Software Tools.

Stimuli

The faces used in this study were from the AR database (Martinez & Benavente, 1998). This database contains full-color photographs of over 100 persons under various conditions. In an attempt to eliminate obvious recognition cues, we only included photographs of 17 men without facial hair or eyeglasses and with any distinctive moles or acne removed using Adobe® Photoshop. Only images of men were used because women's hairstyles and use of cosmetics are often quite distinctive.

The images were pre-processed to simplify subsequent analyses. First, the photographs were converted from RGB to grayscale images. Then the images were normalized, first for orientation by rotating the image such that the eyes share the same vertical position, then for scale by resampling the images to align the mouths, chins, and ears and to fill a frame of 156 $\times$ 215 pixels (Martinez, 2003). Finally, to increase contrast and to give all the images the same dynamic range, the histogram equalization algorithm in MATLAB® Image Processing Toolbox was applied. Six different images resulting from these normalization procedures are shown in Figure 1. Note that these faces have four possible expressions (neutral, angry, smiling, and screaming) and three possible patterns of illumination (ambient, spotlight on the left, and spotlights from both the left and right).

Procedure

Observers performed a match-to-sample task as outlined in Figure 2. A neutral expression under ambient illumination was used as the sample face on every trial. This was followed by two alternatives, which had changed in either expression or illumination. One alternative shared the same identity as the sample (the "match"), whereas the other did not (the "foil"). The observers were instructed to ignore any changes in expression or illumination and to select which of the two alternatives depicted the same person as the sample. Each trial began with a fixation cross for 2000 ms. Then the sample face was presented for 650 ms, followed by a 500 ms mask consisting of a random grouping of textured objects. The alternatives were then displayed for 650 ms, again followed by a 500 ms mask. These presentation speeds were selected based on pilot experiments to avoid ceiling and floor effects, such that the overall level of accuracy would be approximately 75%. Observers made a key press response to indicate which of the two alternatives matched the identity of the sample. If no response was detected before the next sample face was presented, the trial was excluded from subsequent analyses. Before the experiment began, there was a practice sequence of ten trials with feedback. For the experiment itself, no feedback was given. A total of 200 trials were presented in 10 blocks of 20 trials with short breaks in between blocks.

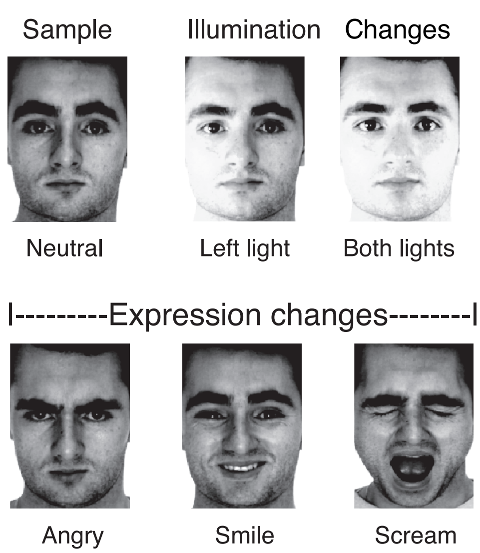

Figure 1. Conditions from the AR database used in the experiments. These images have been converted to intensity images and warped as described in Martinez (2003).

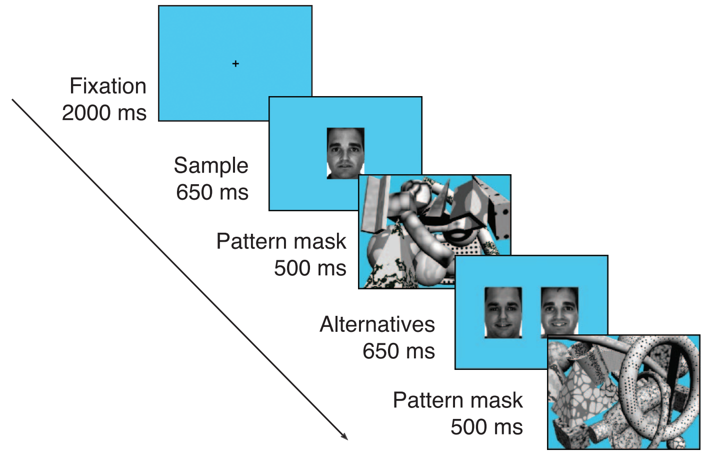

Figure 2. Trial sequence for Experiment 1.

Trial construction

The selection of images employed in this experiment was designed specifically to make many of the identity judgments difficult for appearance-based models. For example, if the similarity between the sample and match images was always greater than the similarity between the sample and foil images, then any appearance-based measure would perform at or near 100 percent accuracy. In order to prevent this, the stimulus set was constrained so that the range of differences between the match and the sample, as measured by correlation, would be approximately equal to the range of differences between the foil and sample. Although image correlation is only one of many possible measures that could be used to constrain the construction of stimulus triads, this procedure ensured that appearance-based models would produce incorrect responses on a substantial number of trials.

Observers

Twenty nine Ohio State University students participated in the experiment; 18 received course credit and 11 were paid. All had normal or corrected-to-normal vision.

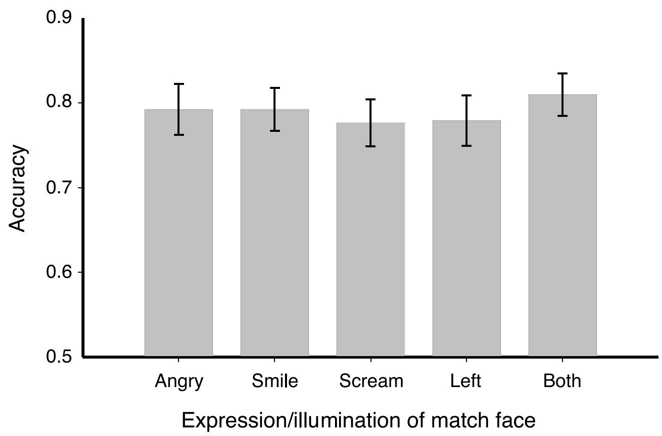

Figure 3. Observer accuracy across expression and illumination conditions in Experiment 1.

Results

Figure 3 shows the percentage of correct responses for each of the illumination and expression conditions, excluding the 0.0039 proportion of trials for which no response was recorded. Overall, the mean level of accuracy was 79% and the average response time was 938 ms. Note that there were no significant differences in performance among any of pair-wise combinations of conditions involving changes in facial expression or the pattern of illumination.

To assess the predictions of appearance-based models for performing this task, we used several commonly employed techniques for representing low level image structure. One approach is to consider each image as a vector in a high-dimensional space, where each individual pixel defines a dimension, and the intensity of the pixel defines a specific position along that dimension. An alternative representation that is perhaps more biologically plausible is to approximate the encoding of image structure as a set of filter outputs that are designed to mimic the responses of simple cells in area V1 of the visual cortex. In our implementation of this approach, these cells were modeled as log Gabor filters with six different orientations with a separation and bandwidth of 30 degrees, five different scales with a separation and bandwidth of 1.4 octaves, and two different phases (even and odd symmetric) in all possible combinations. The selection of five scales was constrained so that the wavelength of the smallest filter would cover at least three pixels, and the wavelength of the largest filter would be no larger than the size of the image. The output of each filter at each image location was computed in the Fourier domain as described by Kovesi (1999). Much like a pixel-based representation, the set of filter outputs for any given image can be thought of as a vector in a high-dimensional space, where each individual Gabor filter defines a dimension, and the output of the filter defines a specific position along that dimension.

We also performed a principal components analysis in order to produce more streamlined versions of both the pixel and Gabor representations. One hundred principal components were extracted from an independent training set of 858 images of male faces from the AR database (Martinez & Benavente, 1998). The training and test sets were mutually exclusive to mimic the novelty of the faces for the observers in our experiment. The training set included images of 34 individuals with all of the expression and illumination conditions used in the present experiment, plus several others that were not used in order to better simulate our subjects' breadth of experience with faces. These additional conditions included illumination with a bright spotlight from the right and images of individuals who wore sunglasses or a scarf.

For the PCA representations, each image was approximated as a linear weighted sum of the principal components (i.e., eigenfaces) that were calculated from the covariance matrix obtained from the training set of images (Turk & Pentland, 1991). As with the pixel and Gabor representations, an image is again considered as a vector in a high-dimensional space, but the dimensions are defined by the principal component weights rather than pixel intensities or the outputs of Gabor filters. In the PCA representation, we employed a total of 100 components, which accounted for over 99% of the data variance. Following Pentland et al. (1993), we excluded the first three principal components from the representation, because they are often most heavily influenced by variations in illumination as has been demonstrated by Belhumeur and Kriegman (1998). This produced slightly improved fits of the PCA models to human performance in the present experiment than when the first three components were included.

For each of these alternative representations, we used multiple metrics for quantitatively measuring the difference between any pair of image vectors. The first of these measures involved computing the Euclidean distance between their respective vector endpoints. We also performed a dot product on each pair of image vectors to compute the angle between them. The primary difference between these approaches is that the distance measure is sensitive to variations in image contrast, whereas the angle measure is not. For the PCA representations, we also employed a Mahalanobis distance metric, in which the space is warped according to the variances and covariances determined by the training samples. This is because PCA will select those dimensions that carry most of the covariance of the data. Using these as a distance metric ensures we appropriately weight the PCA dimensions by the relative amounts of covariance they account for.

One possible method for assessing the psychological validity of face recognition models is to compare their relative accuracy on face matching tasks with the

Measure	Euclidean	Cosine	Mahalanobis
Raw pixels	0.68	0.72	–
Gabor filter outputs	0.71	0.78	–
Pixel PCA	0.74	0.72	0.72
Gabor filter PCA	0.68	0.72	0.76

Table 1. Mean proportion correct for appearance-based measures for Experiment 1. For purposes of comparison, the proportion of correct responses for human observers was 0.78.

performance of human observers (Valentin et al., 1997; Wallraven et al., 2005). To facilitate that analysis in the present experiment we computed the predicted response on each trial for each of the possible difference measures described above. The predicted response in this context between the match and the foil is the one that is quantitatively most similar to the standard. Table 1 shows the percentage of correct responses predicted by each measure. Note that all of the models performed well above chance and that their overall levels of accuracy were quite similar to the performance of human observers with the same stimuli.

A more stringent analysis for assessing the psychological validity of these models is to compare the relative image differences on individual stimulus triads (i.e., the standard, match, and foil) with the accuracy of observers' judgments on those triads. To facilitate that analysis in the present study, we computed the difference in image structure between the sample and match on each trial, and subtracted that from the difference between the sample and foil. Thus, positive numbers would be obtained for trials in which the match was most similar to the sample, and negative numbers would be obtained when the foil was more similar. These difference measures could then be correlated using logistic regression with the percentage of trials that the observers correctly identified the match for each individual stimulus triad. The basic idea from the perspective of an appearance-based model is that observers should be most accurate for triads with large positive difference measures and least accurate for triads with large negative difference measures. The results of these regression analyses are presented in Table 2. Although most of these correlations were statistically significant because of the large number of degrees of freedom, none of the measures could account for more than 20% of the variance in the accuracy of observers' judgments among different triads. For the PCA representations, we also performed a moving window procedure as described by O'Toole, Abdi, Deffenbacher, and Valentin (1993) to see if the fits could be improved by only considering subsets of the principal components for measuring image differences. Although the optimal subsets produced somewhat better fits than the overall PCA, none of them produced r^2 values above 0.22. Thus, these findings indicate that the pattern of errors for these appearance-based models had relatively little overlap with the errors produced by human observers.

In order to interpret these results, it is first necessary to measure the consistency among different observers in their overall patterns of errors. Suppose, for example, that each observer employed a different strategy for performing the required task. If the patterns of errors produced by these strategies were sufficiently heterogeneous, then the lack of regularity in the behavioral data would make it impossible for any model to account for a high proportion of the variance. In order to assess this issue, we employed a modified *K*-folds cross-validation procedure to compare the patterns of errors among different observers (Efron & Tibshirani, 1993). The observers were divided into two near equal subsamples, and we calculated the percentage of correct responses within each subsample for each of the different stimulus triads that were presented over the course of the experiment. The relative accuracies among triads in one subsample were then correlated with those in the second subsample using logistic regression. This was repeated iteratively for all possible subsamples, and then a grand mean r^2 was calculated. The results reveal that there was a high degree of consistency among the different observers such that the average r^2 value from the *K*-folds analysis was 0.715. When considered in combination, these findings provide strong evidence that there was a reliable pattern of errors in the observers' face matching judgments, but that this pattern cannot be explained by any of the appearance-based measures we examined.

In an effort to better understand these results, we sorted all the triads in a spreadsheet based on the difference between human and model performances. The results of this sorting revealed quite clearly that the changes in illumination and changes from a neutral to a scream expression had the largest effects on the image-based models, but that these changes had relatively little impact on the accuracy of observers' judgments.

We also performed an additional analysis to assess any learning that may have taken place over the course of an experimental session. A *t*-test revealed that there was indeed a statistically significant improvement ($p < 0.01$) in the overall accuracy of observers' responses from 76% in the first half of a session to 82% in the second half. This could have resulted from an increased familiarity with the experimental task, or from learning the most salient features of the 17 depicted individuals who were presented over multiple trials with different facial expressions and patterns of illumination.

Measure	Euclidean	Cosine	Mahalanobis
Raw pixels	0.189*	0.039*	–
Gabor filter outputs	0.092*	<0.001	–
Pixel PCA	0.072*	0.089*	<0.001
Gabor filter PCA	0.085*	0.089*	0.156*

Table 2. Logistic regression r^2 values for Experiment 1. Each measure was regressed against mean observer accuracy for each individual stimulus triad. All values with an asterisk are statistically significant ($p < 0.01$).

Results

Figure 6 shows the proportion of different responses for both "same" and "different" trials for each of the checkerboard conditions. Overall, the mean level of accuracy was 72% ($d' = 1.20$) and the average response time was 1098 ms.

The differences between each pair of images were computed using the same computational procedures as described for Experiment 1. For the measures involving PCA, the principal components were obtained from an independent training set of 858 images as described in Experiment 1. Note that these images were not masked by checkerboards.

One obvious strategy for performing same–different judgments within an appearance-based framework would be to set some threshold difference in low level image structure, such that pairs with differences above that threshold would be judged as "different", and all others would be judged as "same". To determine the predicted performance for all of the alternative difference measures, we computed the optimal threshold that would produce the highest levels of accuracy. The results of this analysis for each of the different measures are presented in Table 3.

Note that the PCA measures consistently outperformed those in which the images were represented in terms of pixel intensities or Gabor filter outputs. The highest level of performance was obtained for the pixel-based PCA representation when image differences were computed using the angle measure. Using the most optimal threshold, this measure discriminated faces correctly on 78% of trials ($d' = 1.19$). The results for this measure in all of the checkerboard conditions are plotted in Figure 7.

To measure the consistency of performance across different observers, we used the same *K*-folds cross-validation procedure as described for Experiment 1. The observers were divided into two equal subsamples, and we calculated the percentage of "different" responses within each subsample for each of the possible image pairs that were presented over the course of the experiment. The relative proportions of "different" responses among the stimulus pairs in one subsample were then correlated with those in the second subsample using logistic regression. This was repeated iteratively for all possible subsamples, and then a grand mean r^2 was calculated. The results reveal that there was a high degree of consistency among the different observers such that the average r^2 value from the *K*-folds analysis was 0.723.

Additional analyses were performed to determine if any of the appearance-based models could account for variations in difficulty among the different stimulus pairs. This was achieved by correlating the proportion of "different" judgments for each pair with the magnitude of their low level image differences using logistic regression. The results of this analysis are presented in Table 4 for all of the possible distance metrics we

Measure	Euclidean	Cosine	Mahalanobis
Raw pixels	0.57	0.57	–
Gabor filter outputs	0.65	0.62	–
Pixel PCA	0.69	0.78	0.69
Gabor filter PCA	0.70	0.69	0.69

Table 3. The mean proportion of correct responses for appearance-based measures in Experiment 2, which were calculated for the threshold yielding the highest d' value. For purposes of comparison, the proportion of correct responses for human observers was 0.72.

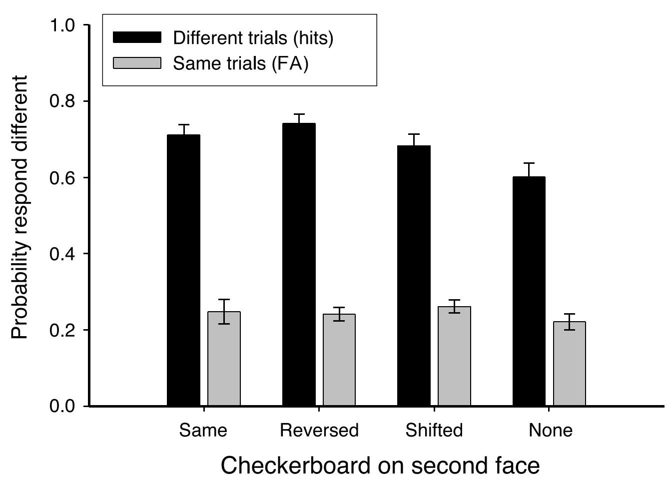

Figure 6. Mean observer performance for all of the checkerboard conditions of Experiment 2.

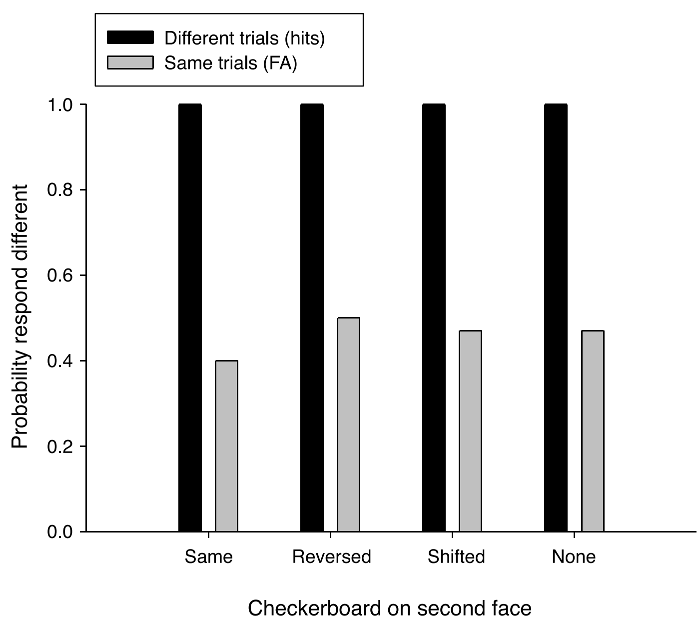

Figure 7. Probability of different responses for the PCA pixel model across all the checkerboard conditions in Experiment 2, using the threshold that produced the highest discrimination performance.

Figure 8. A grayscale image of a face (left), the same image with a randomly scrambled amplitude spectrum (middle), and the same image with a randomly scrambled phase spectrum (right).

study included pairs of complementary grayscale images in which every other Fourier component (8 scales $\times$ 8 orientations) was included in one member and the remaining components were included in the other. Observers were more accurate and had faster reaction times at identifying famous faces or objects when the images were identical to those presented in an earlier block. This priming effect also occurred for objects when a complimentary image was presented, but no priming occurred for complimentary images of faces. Biederman and Kalocsai concluded from this that the representation of a face, unlike that of objects, is specific to the original filter outputs of its Fourier components.

In order to assess the validity of this conclusion it is useful to consider the set of images presented in Figure 8. The image on the left is one of the stimuli from the present experiments. The image in the middle has a phase spectrum that is identical to the one on the left and an amplitude spectrum that was selected randomly from a uniform distribution. The image on the right, in contrast, has an amplitude spectrum that is identical to the one on the left, and a phase spectrum that was selected randomly from a uniform distribution. Note that the information about facial identity is preserved in the middle image even though the amplitudes of all the Fourier components have been randomly scrambled from the original image. This demonstration suggests that it is the alignments of the Fourier components that provide the primary information for the perceptual analysis of faces rather than their amplitudes. Scrambling the amplitude spectrum removes most of the luminance gradients within the original image, but it does not affect the contour structure of the facial features or the polarity of light and dark regions. These are the properties encoded by the phase spectrum that we suspect are most important for the perceptual analysis of faces.

One potentially important difference between the experiments reported by Biederman and Kalocsai (1997) and those reported here is that they used a name verification task with images of famous people, whereas we used a same–different identity task with images of unfamiliar individuals. Hancock, Bruce, and Burton (2000) have argued that the perceptual processing of familiar and unfamiliar faces may be quite different, but that it is the representations of unfamiliar faces that are most likely to be based on relatively low level image descriptions, such as the one proposed by Biederman and Kalocsai. Because familiar individuals have been seen in so many different contexts, it would be reasonable to expect that the representation of their faces would incorporate whatever context invariant properties makes them perceptually distinct from one another. Indeed, this view is supported by the finding that caricatures of familiar faces, which exaggerate distinctive features, are sometimes easier to recognize than the undistorted faces themselves (e.g., see Rhodes, Byatt, Tremewan, & Kennedy, 1997).

Given that the pattern of performance of appearance-based models in the present investigation was quite different to that of humans, it is tempting to conclude that observers may incorporate a feature-based approach for performing same–different identity judgments, or perhaps some hybrid model that combines both approaches (e.g., Schwaninger, Wallraven, & Bülthoff, 2004; Wallraven et al., 2005). The computational analysis of facial features typically involves a graph representation that captures the spatial arrangements of fiducial points, such as the corners of the mouth and eyes (Wiskott, Fellous, Krüger, & von der Malsburg, 1997). The primary limitation of these analyses as models of human face recognition is that there are no reliable procedures for localizing the fiducial points without manual intervention. One attempted solution to this problem is to use corner detectors to localize features in regions that exhibit high curvatures of pixel intensities within their neighborhood (Schwaninger et al., 2004). Although this avoids the need for manual intervention, the fiducial points detected by this procedure are only loosely coupled to those that are marked by human observers. More accurate methods of extracting fiducial points have been developed that use manually marked images to train a system, which then operates autonomously subsequent to this training (Ding & Martinez, 2008; Heisele, Serre,

Cooper, E. E., & Wojan, T. J. (2000). Differences in the coding of spatial relations in face identification and basic-level object recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 470–488. [PubMed]

Craw, I., Ellis, H., & Lishman, J. R. (1987). Automatic extraction of face features. *Pattern Recognition Letters, 5,* 183–187.

Davies, G. M., Ellis, H., & Shepherd, J. (1978). Face identification: The influence of delay upon accuracy of Photofit construction. *Journal of Police Science and Administration, 76,* 35–42.

Ding, L., & Martinez, A. M. (2008). Precise detailed detection of faces and facial features. *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*. Anchorage, AK.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8,* 431–439. [PubMed]

Goldstein, A. J., Harmon, L. D., & Lesk, A. B. (1971). Identification of human faces. *Proceedings of the IEEE, 59,* 748.

Hancock, P. J., Bruce, V. V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4,* 330–337. [PubMed]

Heisele, B., Serre, T., Pontil, M., & Poggio, T. (2001). Component-based face detection. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 657–662). Kauai, HI: IEEE Computer Society Press.

Hill, H., & Bruce, V. (1991). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 263–266.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 986–1004. [PubMed]

Hill, H., Schyns, P. G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62,* 201–222. [PubMed]

Jiang, X., Rosen, E., Zeffiro, T., VanMeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fMRI and behavioral techniques. *Neuron, 50,* 159–172. [PubMed] [Article]

Kanade, T. (1973). *Picture processing system by computer complex and recognition of human faces* (Doctoral dissertation, Kyoto University, Kyoto, Japan). Retrieved from http://www.ri.cmu.edu/pub_files/pub3/kanade_takeo_1973_1/kanade_takeo_1973_1.pdf.

Kaya, Y., & Kobayashi, K. (1972). A basic study on human face recognition. In S. Wanatabe (Ed.), *Frontiers of pattern recognition* (p. 265). New York: Academic Press.

Kovesi, P. D. (1999). Image features from phase congruency. *Journal of Computer Vision Research, 1,* 2–26.

Liu, C. H., & Chaudhuri, A. (2002). Reassessing the 3/4 view effect in face recognition. *Cognition, 83,* 31–48. [PubMed]

Martinez, A. M. (2003). Matching expression variant faces. *Vision Research, 43,* 1047–1060. [PubMed]

Martinez, A. M., & Benavente, R. (1998). The AR face database. *CVC Technical Report #24*.

Meytlis, M., & Sirovich, L. (2007). On the dimensionality of face space. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29,* 1262–1267. [PubMed]

Nachson, I., Moscovitch, M., & Umilta, C. (1995). The contribution of external and internal features to the matching of unfamiliar faces. *Psychological Research, 58,* 31–37. [PubMed]

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

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research, 38,* 2351–2363. [PubMed]

Pentland, A., Starner, T., Etcoff, N., Masoiu, N., Oliyide, O., & Turk, M. (1993). Experiments with eigenfaces. In R. Bajcsy (Ed.), *Proceedings of the 13th International Joint Conference on Artificial Intelligence, Looking at People Workshop*. Chambéry, France: Morgan Kaufman.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19,* 473–497. [PubMed]

Rhodes, G., Byatt, G., Tremewan, T., & Kennedy, A. (1997). Facial distinctiveness and the power of caricatures. *Perception, 26,* 207–223. [PubMed]

Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception, 32,* 285–293. [PubMed]

Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2004). Computational modeling of face recognition based on psychophysical experiments. *Swiss Journal of Psychology, 63,* 207–215.

Sinha, P. (2002a). Qualitative representations for recognition. In *Lecture notes in computer science* (vol. 2525, pp. 249–262). Heidelberg, Germany: Springer-Verlag.

Sinha, P. (2002b). Recognizing complex patterns. *Nature Neuroscience, 5,* 1093–1097. [PubMed]

Sinha, P., & Poggio, T. (1996). I think I know that face… *Nature, 384,* 404. [PubMed]

Sinha, P., & Poggio, T. (2002). 'United' we stand. *Perception, 31,* 133. [PubMed]

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

Tarr, M. J., & Gauthier, I. (1998). Do viewpoint-dependent mechanisms generalize across members of a class? *Cognition, 67,* 73–110. [PubMed]

Tarr, M. J., Kersten, D., & Bülthoff, H. H. (1998). Why the visual recognition system might encode the effects of illumination. *Vision Research, 38,* 2259–2275. [PubMed]

Troje, N. F., & Bülthoff, H. H. (1998). How is bilateral symmetry of human faces used for recognition of novel views? *Vision Research, 38,* 79–89. [PubMed]

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

Valentin, D., Abdi, H., Edelman, B., & O'Toole, A. J. (1997). Principal component and neural network analyses of face images: What can be generalized in gender classification? *Journal of Mathematical Psychology, 41,* 398–413. [PubMed]

Wallraven, C., Schwaninger, A., & Bülthoff, H. H. (2005). Learning from humans: Computational modeling of face recognition. *Network, 16,* 401–418. [PubMed]

Wiskott, L., Fellous, J. M., Krüger, N., & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19,* 775–779.

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14,* 737–746. [PubMed]