A comparison of two computer-based face identification systems with human perceptions of faces.

Peter J.B.Hancock and Vicki Bruce
Department of Psychology
University of Stirling
Stirling
FK9 4LA
Scotland

**November 1997: To appear in Vision Research**

A. Mike Burton
Department of Psychology
University of Glasgow
Glasgow
Scotland

## Abstract

The performance of two different computer systems for representing faces was compared with human ratings of similarity and distinctiveness, and human memory performance, on a specific set of face images. The systems compared were a graph-matching system (e.g. Lades et al., 1993) and coding based on Principal Components Analysis (PCA) of image pixels (e.g. Turk & Pentland, 1991). Replicating other work, the PCA-based system produced very much better performance at recognising faces, and higher correlations with human performance with the same images, when the images were initially standardised using a morphing procedure and separate analysis of "shape" and "shape-free" components then combined. Both the graph-matching and (shape + shape-free) PCA systems were equally able to recognise faces shown with changed expressions, both provided reasonable correlations with human ratings and memory data, and there were also correlations between the facial similarities recorded by each of the computer models. However, comparisons with human similarity ratings of faces with and without the hair visible, and prediction of memory performance with and without alteration in face expressions, suggested that the graph-matching system was better at capturing aspects of the appearance of the face, while the PCA-based system seemed better at capturing aspects of the appearance of specific images of faces.

## Introduction

Engineers attempting to build computer systems for the recognition and matching of faces have at their disposal a bewildering variety of possible techniques. For example, faces may be represented and compared as lists of measurements of features (e.g. interocular distance, length of nose, width of mouth)[1], or more complex area or ratio

[1] Some of these measurements, however, would require reliable and automatic methods for locating features, itself a far from trivial problem.

measures (e.g. ratio of face width to face height), or they may be described in terms of parametric variations on an underlying 3D surface (cf. Bruce et al., 1993). However, the techniques which have had most recent success have been based upon relatively low level image-features (note the hyphen: we reserve the term “features” here for facial landmarks such as nose and eyes). For example, in a recent competition funded by the Army Research Laboratory to find the most robust face recognition method (Phillips, Rauss and Der 1996), the systems which reached the final stages were those of Pentland et al, (Pentland, Moghaddam and Starber 1994), based upon Principal Components Analysis (PCA) of image pixel values, Malsburg et al (Wiskott, Fellous, Krüger and von der Malsburg, 1995) based upon graph-matching of Gabor wavelets and Atick (Penev and Atick 1996), based upon Local Feature Analysis, which is derived from PCA.

The apparent success of these systems is particularly interesting given that they all seem to rely not on abstracted information about faces *per se*, but instead code images of faces in terms only of lower-level characteristics. PCA is a method of dimensional reduction which codes statistical regularity in a set of images (in this case faces). Treating each image as a vector of (pixel intensity) values, it is possible to derive eigenvectors, relatively few of which capture most of the variance in the image set. Images of faces can subsequently be represented not as a vector of pixels, but as the weighted sum of eigenvectors which best reconstructs the original. In this way, facial images are represented in terms of the statistical regularities within a set. In contrast, the system based on graph-matching of Gabor wavelets does not capture regularity within a set, but specific image characteristics. The images are filtered by sets of Gabor filters at several scales and orientations (see Figure 4), centred at each of a number of points around the face. These points form a labelled graph, with vertices represented by the outputs of the filters. In the graph-matching system described by Wiskott et al (1995), the graph derived from a particular facial image is matched against a set of stored graphs derived from other face images. Successful matching occurs as a function of goodness of fit measures derived from these matches.

Although these classes of system are very different from one another, they share the characteristic of being based on image properties, rather than abstract representations of faces. Interestingly, there is now considerable evidence that human representation of faces may similarly be based upon relatively low level image-features rather than more abstract descriptions of facial features (i.e. feature separations or protuberances). Suggestive evidence for this proposal comes from the fact that human face recognition is severely disrupted by certain image transformations which might be thought to leave abstract representations untouched. First, human face recognition is considerably disturbed if faces are shown in photographic negatives (e.g. Galper and Hochberg, 1971; Bruce & Langton, 1994), and it is not clear why this should be so if simple feature measurements formed the basis of our representations. Recognition and matching of faces is also severely disrupted by changes in direction of lighting, particularly where unfamiliar lighting directions are used (e.g. Hill & Bruce, 1996; Johnston, Hill & Carman, 1992). Human face recognition is also very poor if simple line drawings are shown which depict the outline of face features (thereby preserving their spatial layout), but very much enhanced if a simple thresholding operation is added which blacks in areas which were originally dark in the image (Davies, Ellis & Shepherd, 1978; Bruce et al, 1992). Again, it is not clear why this result should occur

if faces were represented as lists of feature measurements, although effects of negation, image thresholding and lighting might all be explicable if 3D shape-from shading were an important component of building face representations. However, if human face perception delivers a 3D model of a face it is unclear why recognition of previously unfamiliar faces is so badly disrupted when viewpoint is changed (e.g. Bruce, 1982).

Such observations have led us to suggest that the human visual system encodes faces on the basis of an image coding which preserves information about relative light and dark, which may permit recovery of shape from shading. The artificial face recognition systems described by Turk and Pentland (1991) and Malsburg (Lades, Vorbrüggen, Buhmann, Lange, von der Malsburg, Würtz and Konen, 1993, Wiskott et al, 1995) have each been attributed some possible biological plausibility. Principal Components Analysis is readily implemented by neural networks, and Turk and Pentland (1991) speculate that human face recognition is a good candidate for a recognition mechanism based upon fast, low-level 2D image processing, in contrast to the multi-stage model-based approach common in descriptions of 3D object recognition. The development of von der Malsburg et al's system was guided by a dynamic link theory of binding (von der Malsburg, 1985; Bienenstock and Doursat, 1991) and the Gabor wavelets which it uses have been likened to receptive fields in primary visual cortex (e.g. Jones and Palmer, 1987). These systems are described in more detail below.

Given the appeal made by the authors of such systems to their psychological/biological plausibility, there have been rather few attempts to evaluate them against human data. A number of authors have examined how well human recognition memory for faces, or ratings of face distinctiveness, memorability and so forth, correlate with measures of distinctiveness/typicality derived from PCA applied to the same set of images (e.g. O'Toole et al, 1991, 1994; Hancock, Burton & Bruce, 1996). In general, PCA yields good predictions of human performance with the same set of face images. Thus O'Toole et al (1994) found that measures of face distinctiveness delivered by PCA coding cross-loaded with human face ratings and with d' scores obtained from memory experiments. Hancock et al (1996) demonstrated that separate coding and recombination of faces via "shape" and "shape-free" image components (see section on computer analyses, below) produced better performance still. However, one disadvantage of all previous evaluations of PCA against human performance with faces is that it is memory for the pictures, rather than of the faces, which has been tested. As noted by a number of authors (Bruce & Young, 1986, Bruce, 1982; Klatzky & Forrest, 1984), typical face memory experiments in which identical pictures are shown at study and test confound face recognition with picture memory. One aim of the work described in this article is to examine how well PCA is able to predict which faces will be well remembered when they must be recognised in different expressions.

Research which examines how well performance of the graph-matching system correlates with human performance has been more limited still, though Biederman (unpublished) reports some preliminary data using a face matching paradigm. In these studies, the graph-matching system has proved able to predict how human matching of faces declines as viewpoint between the faces-to-be matched is increased.

However, the system has not been used to predict the similarity space between different individual faces.

In none of the above studies has any attempt been made to compare the performance of the system under test with another image-based coding system, so where there are correlations with human performance we do not know if these arise because of general aspects of the system (e.g. coding of image-features) or from something more specific. In the work presented here, we compared the performance of human observers against both graph-matching and PCA systems, to identify which of the two computational models provides a better prediction of human performance, and why.

Three different areas of human performance with faces were examined: similarity; rated distinctiveness; and actual memory. For similarity judgements, observers were asked to sort face images into piles, based on their appearance. For distinctiveness, they were asked to rate how easily they thought they would recognise the face. Subsequently, their actual ability to recognise the faces was tested in a surprise recognition task, thereby producing actual memory scores. The following section describes the collection of these data in more detail. We then describe the two computer-based recognition systems and finally, compare human and computer perceptions of the same faces.

## Human perceptions of faces

### *Materials*

Our image set was derived from 50 young adult Caucasian males. Each person was photographed from a set distance with a neutral expression, and then asked to display happiness, disgust and surprise. However, not all of the original 200 images were usable for technical reasons and in addition, some of the "expressions" were not entirely convincing, leaving us with 186 usable images. We used the set of fifty neutral expressions as targets, and an additional 2 or 3 extra images of each person, a total of 136, as a test set.

### *1. Distinctiveness and memorability measurements*

#### Materials

The fifty neutral faces and fifty picked arbitrarily from the expressing faces, one per identity, were used. They were displayed at 240x280 pixels, corresponding to approximately 5cm width on a Sun monitor, with 256 grey levels. Backgrounds were removed manually using a digital editing program and replaced with a plain white field. Each set was split arbitrarily into two halves, A and B. An example of one of the expressing faces is shown in Figure 1.

#### Participants

60 volunteer white Caucasian undergraduates, approximately half male, were paid to take part in the experiment. The participants did not know any of the people shown in the photographs.

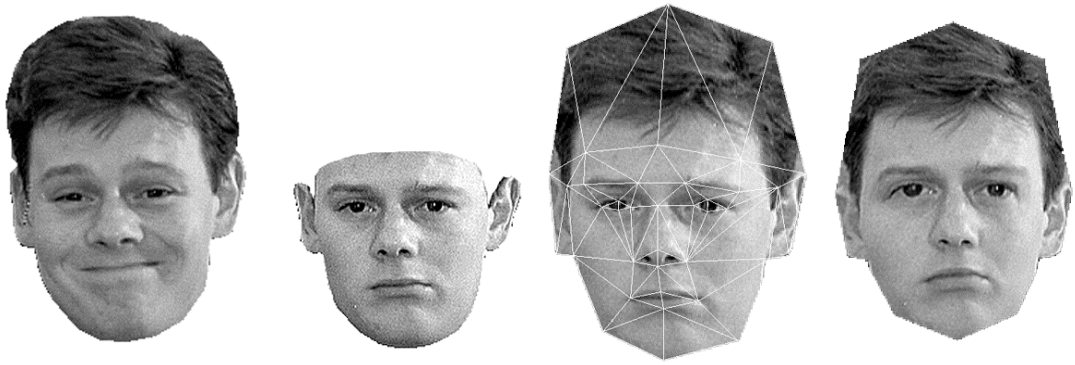

**Figure 1. From left: Original face image, from the expression set; neutral version of same face with hair manually removed; triangulation used to mark the "shape" of the face; "shape free" face after morphing to the average shape.**

## Method

Participants were initially shown 25 faces, from one of the half sets A or B, sequentially and asked to rate each for distinctiveness on a scale of 1-10, prompted by the question: "suppose you had to meet this person at a station, how easy would it be to pick them out". A score of 10 corresponds to easy, meaning distinctive.

After a gap of 10-15 minutes, while they were doing unrelated object-recognition tasks, participants were shown fifty faces, without prior warning. They were asked to respond, again on a scale of 1 to 10, how certain they were that each face had been shown in the first part of the experiment, with 10 being certain that it had. The responses to those that had in fact been seen were averaged to give a hit score for each face, while those for distracter faces were averaged to give false positive scores.

There were three test conditions. One set of participants were shown neutral faces initially, and subsequently tested on the same set (NN). A second set were shown neutral and tested on expressing (NE), while the third were shown expressing and tested on neutral (EN). In the NN condition, the images used in rating and recognition phases were identical, allowing the possibility of image rather than face recognition. There were 20 participants in each condition, half seeing the A set for rating, half the B set.

The values for distinctiveness, hit score and false positive score from the 10 participants for each half set were combined to give average values for each face. Since two groups (NN and NE) rated the neutral faces for distinctiveness, these values could be averaged over 20 ratings for subsequent comparison with computer data.

## Results

**Table 1: average ratings for the three test conditions.**

| | Distinctiveness | SD | Hit score | SD | False positive | SD |
|---|---|---|---|---|---|---|
| NN | 6.1 | 1.1 | 7.5 | 1.3 | 3.5 | 1.2 |
| NE | 6.1 | 1.1 | 6.5 | 1.5 | 4.5 | 1.1 |
| EN | 6.0 | 0.90 | 6.3 | 1.5 | 4.0 | 1.5 |

| | d’ | SD | criterion | SD |
|---|---|---|---|---|
| NN | 1.9 | 1.2 | 0.12 | 0.56 |
| NE | 0.88 | 1.0 | 0.05 | 0.42 |
| EN | 1.3 | 1.1 | 0.27 | 0.65 |

Table 1 shows mean and standard deviations of the results for the three different conditions. There is no significant difference in the three distinctiveness ratings (Anova, $F(2,98)=0.27$), but there is an effect of condition on memorability of the faces (hit score: $F(2,98)=24.8$, $p<0.01$; false positive score: $F(2,98)=14.3$, $p<0.01$). Our observers did find it easier to recognise the same image than the same face with a different expression. That they should also find it easier to reject distracter images in the NN condition is presumably because of their increased confidence with recognising the correct targets, since the distracter images in all three conditions are equally unseen!

A high hit score for an individual face does not necessarily mean it was memorable: it could be that it also has a high false positive score. In this case, observers are responding yes whether or not it was actually seen before: the face simply seems familiar. This effect has been termed “context free familiarity” (Vokey and Read 1992), since the confusion presumably comes from the participants’ other experience of faces, though it could also be that the face really does resemble one in the distracter set.

By signal detection theory, a high familiarity (= high hit and false positive) score would correspond to participants using a low criterion. Highly memorable faces, with high hit score and low false positive score, will have a high value for d’. Although it is not usual to compute d’ values using single observations from multiple observers, we can do so. The responses are binarised such that responses of 6 or above are counted as “yes”, 5 and below as “no”. These average hit rates yield values for d’ and criterion, average values for which are reported in Table 1. The values for d’ confirm an effect of condition on memorability ($F(2,98) = 26.3$, $p<0.01$) but there is no effect on criterion ($F(2,98) = 2.86$, $p>0.01$).

### *2. Similarity ratings*

A sorting task was used to gather information about the perceived similarity of our set of faces. The task was carried out with and without hair information.

#### Materials

The same 50 male face images, on a white background as before, were printed at 3.5x4.5 inches on a laser printer and enclosed in clear sleeves to prevent handling damage. A second set of prints was produced following editing to remove the hair from the images, see Figure 1.

#### Participants

80 volunteer white Caucasian undergraduates, approximately half male and none of whom had participated in experiment 1, were paid to take part in the experiment, half sorting images with hair, half without.

### Method

The participants were given the set of 50 prints, and asked to sort them into sets based on facial similarity. They were free to choose how many subsets to create and what the basis for separation should be. If prints are laid on top of each other in piles, there is a temptation to match only to the most recent member of a pile. Participants were therefore encouraged to spread them out on a large table or the floor, in order to help with the identification of coherent groupings. The process typically took about twenty minutes.

### Results

Participants sorted the faces with hair into an average of 11.5 piles, ranging between 2 and 29. Without hair, they were sorted into an average of 9.8 piles, ranging between 2 and 23. The somewhat lower numbers without hair reflect the expected extra difficulty of the task.

The most direct way to assess similarity from these sort data is simply to count the number of co-occurrences: the more often two faces are grouped together, the more similar they are taken to be. The range of co-occurrences with hair is from 0 to 20, with a mean (and median) of 6, without hair, from 0 to 22, again with a mean of 6.

## Computer analyses

### *Principal component analysis*

PCA is established as a method for computer identification of faces (Turk and Pentland 1991). Analysis of a set of suitably aligned face images yields a set of "eigenfaces" which may be used as the basis for a compact coding of the faces (see Figure 2). New images are analysed using the same set of eigenfaces to give a vector which may be matched rapidly using Euclidean or some other measure of distance with the stored codings. In the work reported here, the images are normalised for inter-ocular distance and ocular location, i.e. the faces are scaled and translated to put the centre of both eyes in the same x,y location for all images.

Craw and Cameron (1991) showed that recognition can be improved if the faces are first "morphed" to an average shape prior to running PCA. A number of points, in this case 38, that define the locations of features such as eyes, nose and mouth are located on each face manually (see Figure 1). It would be possible to do this automatically, as the graph matching system below does, but we have not implemented such a system. The morphing procedure then aligns all the major internal features of each face, putting them into an average position. PCA may then be performed separately on the shape-free face images and on the shape vectors consisting of the x,y location of the points on the original face images. The shape components capture much of the effect of changing expressions, as well as the variations in shape between individuals, while the shape-free image components capture fine scale detail such as variations in nose shape and skin texture.

**Figure 2 The first four "shape free" eigenfaces.**

The variations captured by the shape components may be viewed by creating an animation. An average face is distorted by displacing the key points by a range of small amounts (e.g. +/- 0.5 in steps of 0.1) in the direction specified by a shape component. When viewed sequentially, the first component in our set is seen to produce a nodding head. Note that while all the faces are supposedly looking straight at the camera it is not possible to control this completely. The first component thus codes the angle of the head, which seems unlikely, except perhaps in extreme cases, to have much bearing on human perceptions of distinctiveness or identity of the face. The second component codes head size, which presumably is relevant. The third codes the position of the major feature groups within the face, but the fourth and fifth code the other two dimensions of head movement: shaking and twisting side to side. These two components, along with the first, were removed from subsequent computations on the grounds that while of relatively high variance, they should contribute little to any possible human perceptions of important aspects of the face. Animations showing the effects of these components are available at http://www-psych.stir.ac.uk/~pjh/pca-face.html.

One issue that always arises when using PCA is how many components to take. We have 50 faces, which will give up to 49 components, after subtracting the mean. O'Toole, Abdi, Deffenbacher and Valentin (1993) showed that later components (those with lower eigenvalues) were more important than the earlier components for recognition. Early components carry more general characteristics of the faces, such as their gender. Hancock et al (1996) showed that it is relatively early components that correlate with human perceptions of distinctiveness and memorability. Here we are comparing PCA with human perceptions of distinctiveness and also of similarity and our expectation is that the early components will again be the most significant. To anticipate our results somewhat, Figure 3 shows correlations with similarity and distinctiveness as the number of components is varied. The details of what is being plotted are explained below, for now we are only interested in the general form of the graphs. Much of the correlation comes from the very first component and in both cases rises to a maximum after only a few components are included. Removing the early components from the correlation causes it to fall away rapidly, leaving only noise by the time ten are excluded.

Similar results are obtained for all the varieties of PCA considered here. Exactly where the maximum occurs varies, but in all cases it was at less than 10 components.

Rather than fiddle about using different numbers of components in each condition, we shall use the first 10 in all our tests, since we are more interested in patterns of results than in the ultimate performance and the fall-off of correlation is very slight in any case. The stability of the plots in Figure 3 indicates that we are unlikely to be missing much.

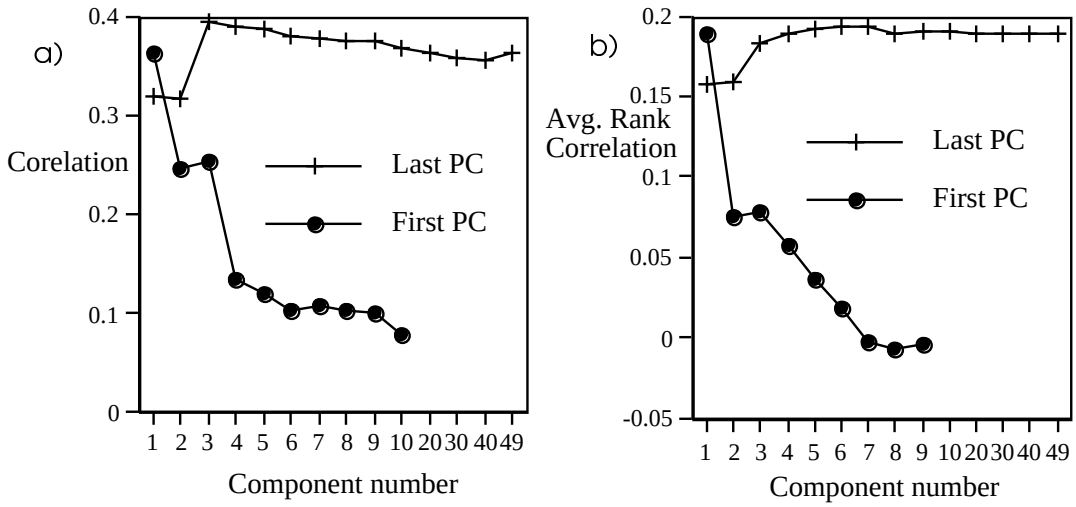


**Figure 3 The effects of varying the number of "shape free" components on correlations with human data. a) Correlation with d' scores b) Average rank correlations with similarity ratings. For each graph, Last PC includes all components from 1 up to that one, First PC includes all components from that one up to 49.**

This leaves the question of how to combine the shape and shape-free image components. Some form of scaling is necessary, since otherwise the shape components will dominate as they have a larger numerical variance, for no interesting reason (one set of components being in the dimension of pixel co-ordinates, the other in the dimension of pixel grey-levels). Since we have no a priori reason to suppose either component type has more significance, the two sets are scaled to have equal variance prior to combination. We take the first 10 shape-free image components, and the first 10 shape components, excluding 1,4 and 5. Plots similar to those in Figure 3 (but not reproduced here) indicate that this combination is in a stable region, where adding or removing one or two components either way has little effect on the results.

We therefore have 4 sets of PCA data for the faces: full image, as per Turk and Pentland, shape-free images, as per Craw and Cameron, the shape vector alone and finally shape + shape-free components combined. The components are generated from the target set of 50 faces, and subsequently used to analyse the test set of 136. All the results here use the first ten components from images (with or without shape removal) and the first 10, minus 1, 4 and 5, from the shape vectors.

### *Graph matching system*

The graph matching face recognition system tested here was developed by von der Malsburg's group at Ruhr-Universitat Bochum (Lades et al, 1993). A full explanation of its operation would be impractical here, and readers are referred to

Lades et al (1993). In summary: face images are coded by families of Gabor-type wavelets (see Figure 4), at several scales and orientations, located at a number of locations around the face. The locations are found automatically for a new face, by comparing the image with a set of reference model faces. Example graphs are shown in Figure 5. The face locations form a labelled graph, with the activity vector (or "jet") of the local wavelets attached to each vertex. A graph is stored for each face in the target set. During recognition a new face image is input and an initial graph formed by reference to the model faces. This is then adapted to form the best possible match with each stored graph in turn, by distorting the grid. The process is controlled by a penalty function which favours similarity of the jets attached to corresponding vertices of the two graphs, and penalises metric deformation of the grid. The test face is deemed to match the stored graph for which the penalty function is lowest. However, it is possible that the input face is unknown to the system, in which case all of the matches should be poor. A measure of confidence in the result is given by the difference between the lowest match distance and the average of the rest. Only if the best match is significantly below the average will an identification be claimed.

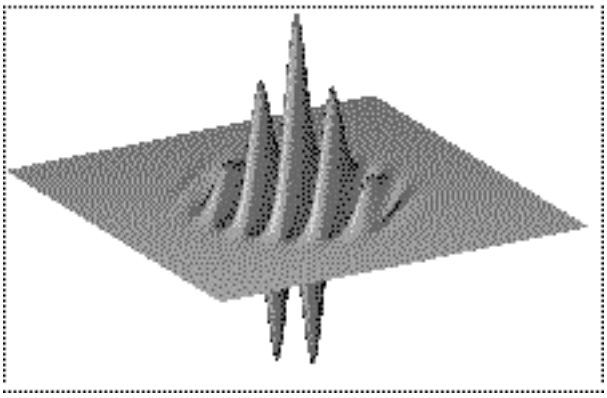

**Figure 4 Response of an example Gabor filter.**

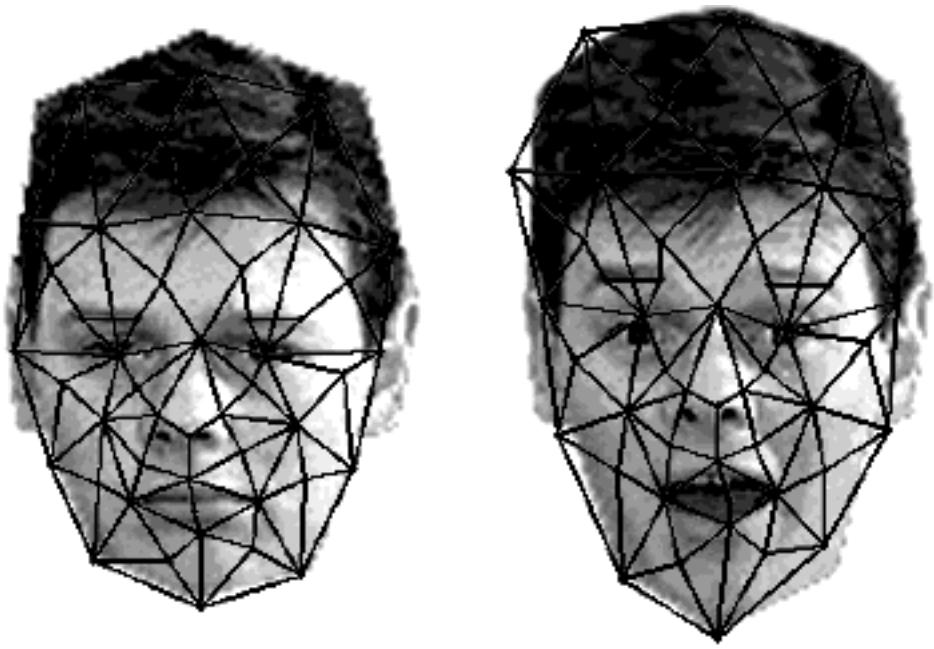

**Figure 5. Example of the graph matching process: target face on the left, test image on the right.**

### *Comparison of the computer systems*

The two systems we are considering differ in an important respect. PCA is a linear transformation of the image space and is purely descriptive. It identifies a set of image-features that capture as much as possible of the pixel information present for a given limited number of descriptors. The graph matching system codes relationships between its low level image-features and is therefore inherently more powerful as a recognition system. This does not mean it will necessarily work better: the power may not be needed, or the image-features may be inappropriate.

PCA on full images is purely image-based. One of the reasons for its limited performance is that facial features will appear in differing locations within the image. The faces are aligned such that the centres of the eyes are in a fixed location within the image but all other features are moveable. The PCA has to code this variability within the eigenfaces, some of which display multiple features. Separating out the coarse shape information allows the image PCA to work in a more linear space - the features are more closely aligned.

In addition to reducing variability between identities, the shape removal can also remove some of the effects of expressions, and provide a measure of invariance to head angle. It can only help with expression effects that change the location of features, such as eyebrows: the sudden presence of teeth in a smile is necessarily an image change. Head angle changes of a few degrees can be allowed for: anything large enough to obscure some features will necessarily pose problems for the image part of the PCA.

The graph matching system as tested here handles both changes of expression and viewing angle similarly, by allowing distortion of the matching grid. Work is in progress to handle rotational invariance more explicitly.

## Computer-human correlations

### *Similarity judgements*

Our human participants gave similarity measures for the target faces by sorting them into piles. These data may be accumulated to give a similarity measure between any pair of faces by counting the occasions on which they are grouped together. The two computer systems also give similarity information. The graph matching system works by explicitly calculating a stress value for the fit between the target face and each face in the recognition set. The lower the stress, the more similar the two faces. The PCA-based system computes a vector for each face. A simple measure of similarity between two faces is then given by the Euclidean distance between their vector representations. This section reports a comparison between the two computer-based similarity metrics, and the human ratings.

One way to approach this is to look at the nearest neighbour for each face: do the computer systems pick as their second choice a face that humans often group with the

target? This approach, however, ignores all the information about the similarity of the other faces in the set. This information might be utilised by performing a correlation between the human and computer similarities for all the faces. A simple correlation seems inappropriate, since the human similarity scale, while hopefully monotonic, seems unlikely to be linear. We therefore performed a rank correlation. For each face in turn as target, the other faces were ordered according to their similarity, as judged by human and computer. A Kendall rank correlation, with adjustment for ties, since there are many in the human data, was then calculated on the two orderings. This gives a total of 50 rank correlations, one per target face, for each comparison between the human and computer data. Average values over the 50 faces are shown in Figure 6. The numbers of components used for the different types of PCA are as described in the section on PCA above. Note that while the human participants were rating the faces with and without hair, the computer systems are only looking at the faces with hair. We wish to establish not what the computer systems would do if presented with a hairless face, but what they do when presented with a normal image with hair, e.g. to what extent might they be "distracted" from the inner features when hair information is present?

Figure 6 shows that all of the numbers are above the line that marks an average correlation of 0.057, which corresponds to $p<0.005$, so that we have formally significant agreement between humans and computer systems. As the the correlations themselves are quite small and the values approximately normally distributed, we perform statistical tests upon them untransformed.

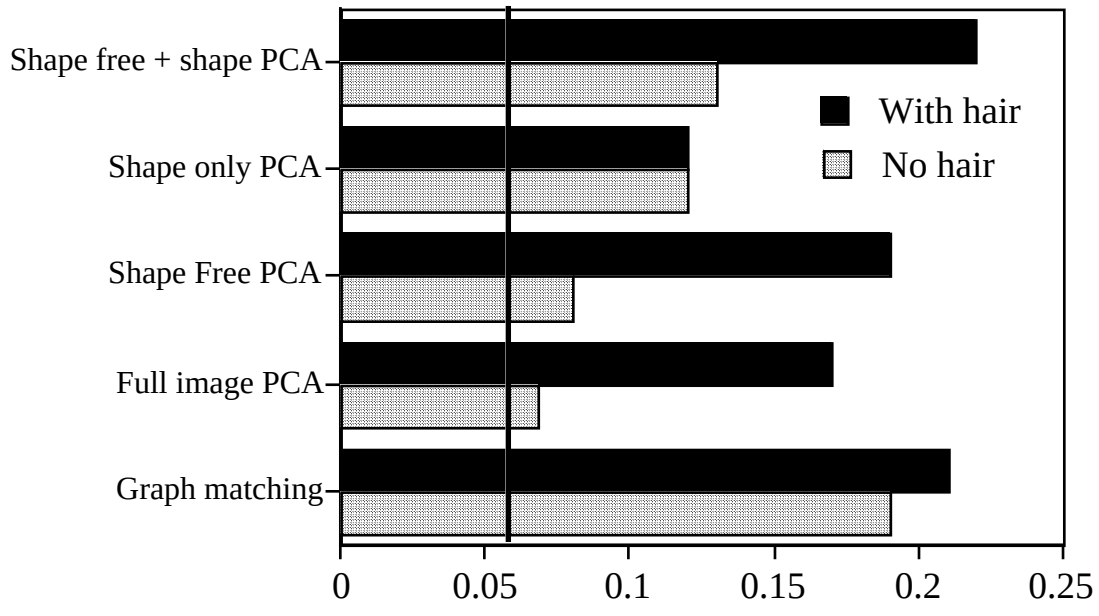


**Figure 6 Average Kendall rank correlations between human and computer-based similarity judgements, when humans are rating faces with and without visible hair. The line marks correlation for $p<0.005$.**

If we consider differences between the systems, two things are apparent from Figure 6: the average orrelations for the graph matching system when compared with humans

seeing faces without hair are higher than for anything from the PCA-based system, and the figures for matching with hair are higher than those without. A paired t-test shows no difference between the shape + shape free PCA and graph matching systems' results with hair ($t_{49}$=0.98, p>0.05). Anova, comparing graph matching with shape + shape-free PCA with and without hair, shows no effect of computer system ($F_{1,49}$=2.9), but an effect of hair presence ($F_{1,49}$=11.1, p<0.005) and an interaction between them ($F_{1,49}$=11.9, p<0.005). It is evident that most of the difference between the systems comes from the no hair condition. The graph matching system appears almost equally able to capture the similarities used by humans, whether their judgements come from faces with or without hair ($t_{49}$=1.16, p>0.05).

Comparing shape + shape-free with full-image PCA, ANOVA confirms an improvement from the shape separation ($F_{1,49}$=12.3, p<0.05) and again shows an effect of the presence of hair ($F_{1,49}$=38.7, p<0.05) but no interaction ($F_{1,49}$=0.46). Shape-only PCA performs identically with or without hair which suggests that people do use the shape of the face to assess similarity whether or not they can see the hair. It does not improve when humans can see the hair because the shape model contains no information about hair. The combination of shape and shape-free image components again fares best of the PCA systems, both with and without hair.

A final point of interest is whether the two computer systems themselves agree on a similarity ordering. That this is the case is suggested by the observation that the rank correlations themselves tend to be correlated. Thus the scores for graph matching and shape + shape free PCA with human with-hair ratings show r=0.65, p< 0.005. Direct comparison of the orderings produced by the two systems confirms this, producing an average rank correlation of 0.48, higher than any of the values in Figure 6. Conversely, comparison of the shape with shape free PCA shows no significant agreement, with an average correlation of only 0.04. Apparently these two varieties of PCA capture different aspects of similarity, which combine to produce a better match with the human data than either alone.

While the correlations are formally significant, they are small. Neither computer system does a particularly good job of explaining the human variance. One possible explanation for this is that the human data are rather noisy. Biederman (1998) reports much better correlations between the graph matching system and human data derived from a psychophysical presentation method. Two faces are presented in rapid succession, and the participant has to respond simply with whether the faces are the same or not. A pilot experiment using some of our faces also gave larger correlations than those from our sorting task, at around 0.35. We shall explore this method further.

Another possibility is that humans use several similarity metrics, of which our computer systems capture only some subset. Previous work relating computer systems to human performance has indicated that the latter may be mediated by a number of factors not directly addressed by our experiments. Thus O'Toole et al (1994) used principal components analysis to recode their human rating data into factors such as memorability and accuracy, the latter yielding particularly good correlation with their computer analysis. We hope that further investigation of this area will help to tease out these complexities.

### *Distinctiveness and memorability*

The human distinctiveness and memorability data came from showing participants one set of face images for rating and subsequently testing on a larger set. The computers were given a similar task, being trained initially on the complete set of 50 neutral faces. For PCA, this means using this set to extract the principal components; for the graph matching system, building a model for each target face. Both systems recognise the target set perfectly, so there is no useful information to be had about distinctiveness from there. The systems were tested on the complete set of 136 expression images. The full PCA system identified all of these correctly, the graph matching system failed on five[2]. Since none of the images were identical to the targets, there was always some error in the matching process, and it is this error that can be related to human concepts of distinctiveness.

We have three sets of human data, from the conditions NN, NE and EN. The distinctiveness data from the first two may usefully be combined, since both groups were rating the same neutral faces. For memorability measures, the closest match to the computer task is given by the NE condition, since then both humans and computer have a neutral target set and are tested on expressing faces.

For the graph matching system, we have a measure of how confident the system is of its match. We would hope to find a positive correlation between this confidence measure and the human-judged distinctiveness and hit rate. For the PCA systems we have two possible measures. One is a confidence measure similar to the graph matching system: what is the difference between the distance to the desired target and the mean distance to all the others, as measured on the axes defined by the principal components. The second is the error of reconstruction: a measure of how well the test image is coded by the PCA basis faces. Previous work has shown that there are correlations between reconstruction error and human ratings of distinctiveness and memorabilty (O'Toole at al 1994, Hancock et al 1996). Here, for consistency with the graph matching system, we report results from the PCA confidence measure, but error of reconstruction showed a very similar pattern of results.

**Distinctiveness**

The results are summarised in Figure 7, which also indicates with a vertical line the value of r=0.24, which is significant at p=0.005 (this low value chosen to adjust for the total of 10 correlations computed).

A number of points are apparent from Figure 7. There is generally little to choose between the correlations with human ratings of expressing and neutral faces. The shape free PCA is better than full image PCA, while the addition of shape information

[2] This should not be taken as evidence that the PCA system is generally better at identification than the graph system. With both systems working at or close to ceiling the diference is not significant, and another set of faces might produce a different result.

again offers some improvement over the shape free image alone. The graph matching system is comparable with the PCA-based system.

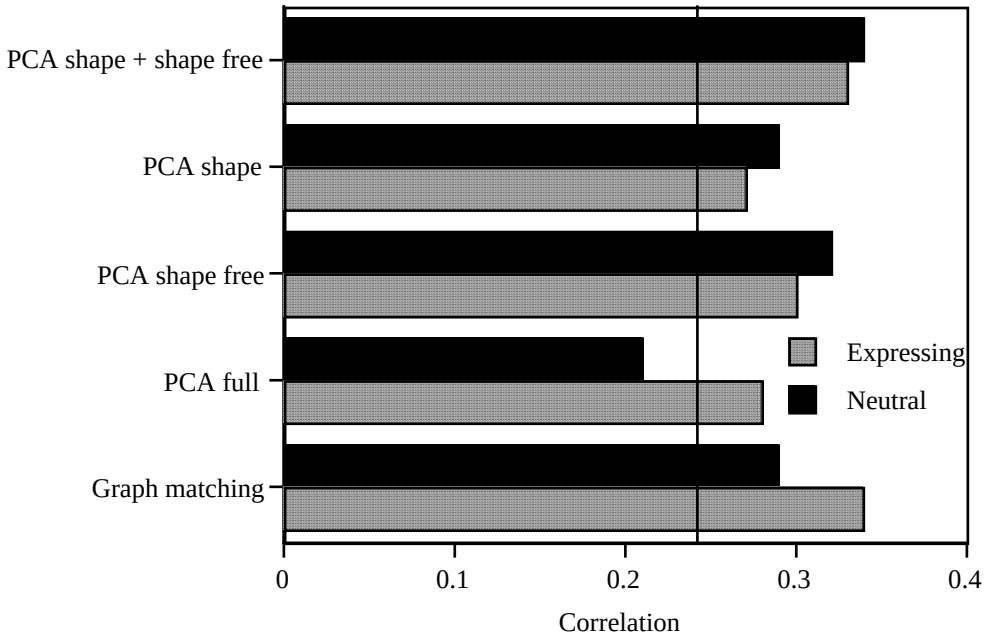


**Figure 7. Correlations between computer systems and human perceptions of distinctiveness.**

## Recognition

Table 2 summarises the correlations with memory scores, showing the performance of each system against hit scores, false positive scores, d' and criterion values. In Table 2 again, $r_{crit}$ for $p<0.005$ is 0.24, so for the NE data, the only significant correlations come from graph matching. For the NN data, many of the correlations are significant. There are no significant correlations with criterion.

| Human NN data | Hit | False positive | D’ | Criterion |
|---|---|---|---|---|
| Graph matching | 0.33 | -0.27 | 0.34 | -0.07 |
| PCA full | 0.30 | -0.16 | 0.19 | -0.07 |
| PCA shape free | 0.42 | -0.36 | 0.37 | 0.04 |
| PCA shape | 0.2 | -0.09 | 0.14 | -0.11 |
| PCA shape + shape free | 0.41 | -0.34 | 0.35 | 0.02 |

| Human NE data | Hit | False positive | D’ | Criterion |
|---|---|---|---|---|
| Graph matching | 0.32 | -0.19 | 0.29 | -0.1 |
| PCA full | 0.15 | -0.07 | 0.1 | -0.03 |
| PCA shape free | 0.2 | -0.23 | 0.22 | 0.06 |
| PCA shape | 0.0 | 0.05 | 0.01 | -0.09 |
| PCA shape + shape free | 0.17 | -0.18 | 0.18 | 0.06 |

**Table 2 Correlations between computer-based measures and human recognition scores.**

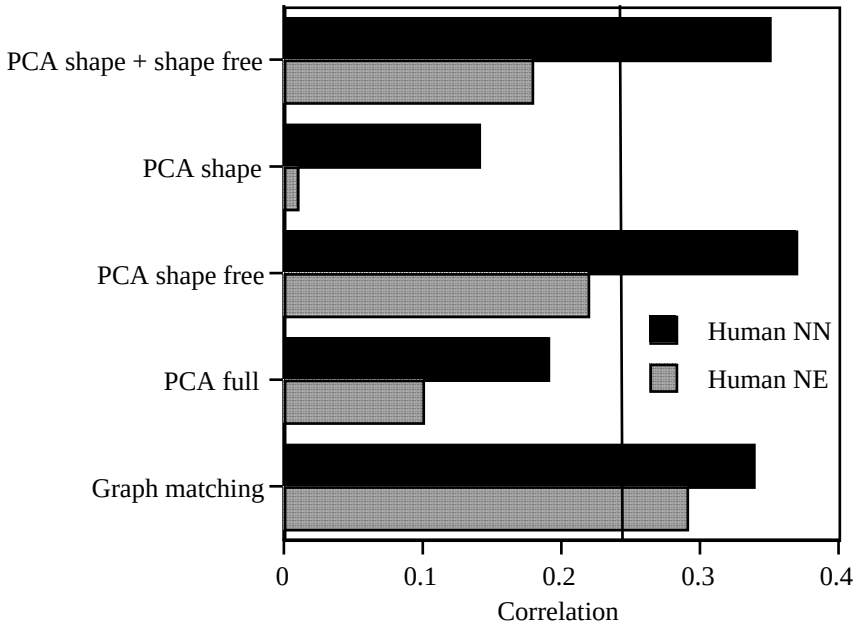


**Figure 8 Correlations between computer systems and human d’ scores.**

Figure 8 shows the results graphically for the derived memorability scores, the others show a similar pattern. Of note is the big difference between the NN and NE conditions for PCA, while the graph matching system shows a much more similar, significant correlation with both. The PCA shape components show little signs of correlation, so it is unsurprising that the combination is similar to the shape-free image components alone.

The correlation shown by PCA for the NN condition, but not for NE, suggests that PCA may be capturing something about the images, rather than the faces, since NN is potentially an image-matching task. To reiterate this point: the computer systems are effectively performing an NE task: trying to identify the expression set of faces by reference to the neutral targets. In the NN condition, humans could simply be matching images, in the NE condition they must to some extent be matching faces. Since PCA , itself doing the NE task, here shows a higher correlation with human performance on the NN task, it suggests that PCA is functioning by matching aspects of the image that stay constant between samples of a face, since that image constancy is presumably what mediates the higher human performance on the NN condition (Table 1).

## Discussion

Examining the performance of our two systems against human assessment of the similarities between different faces; human ratings of distinctiveness, and actual memorability of the faces, shows that both systems are able to capture human performance fairly well, provided that PCA is based upon shape-free faces derived from a morph transformation with shape assessed separately (this replicating the conclusions of Hancock et al, 1996). Moreover, there is some overall correlation between the two systems in terms of the similarities that each sees in our set of faces, as well as between each and human similarity ratings. Despite this, however, detailed analysis of the performance of the two systems shows that where they do well their performance seems to be based on rather different aspects of the appearance of these faces.

The graph matching system seems to capture aspects of the appearance of the faces used rather than individual pictures. The evidence for this is the equivalence of its correlations with human judgements of similarity whether or not the hair was visible to the humans. This suggests that the graph matching system is sensitive to variations in face appearance without being unduly swamped by the hairstyle. Moreover, the correlations between the graph-matching recognition performance and human memory performance remains similar whether humans are doing picture recognition (condition NN) or face recognition (condition NE).

In contrast, the PCA performance seems to be dominated by details of the images. Correlations with human similarities are much higher in the situation where humans judged faces with hair visible, and correlations with memory performance are much better for the picture match condition (NN) than the face match condition (NE). Indeed the correlations achieved by PCA shape + shape-free are better than any achieved by graph-matching when the two pictures match (e.g. correlation of 0.41 c.f. 0.33 on the NN hit rates) but worse than those of the graph-matching system when they do not.

Further evidence that the PCA system is essentially image-based comes from an error observed during the initial phase of this work. The system completely failed to recognise one of the faces and we initially wondered whether it had somehow been mis-labelled. Inspection indicated otherwise, and it took some time to realise that the problem was that this image had accidentally been mirror-reversed during scanning.

This was unnoticeable to the casual human observer, and to the graph-matching system, but the reversal, perhaps because of a slight luminosity gradient, evidently fooled the PCA system.

A question of obvious interest is why the two systems differ in their tendency to look at the face or the image. PCA based on pixels is precise at a pictorial level, once proper linearisation of the image space is achieved via our morphing operations. This shape-free transformation can improve the performance of PCA even if it is working on relatively uninteresting image properties such as contrast gradients, because it will align those in both images. The graph matching system uses Gabor wavelets as local image-feature detectors, which will be relatively insensitive to such image properties. The importance of hair also differs between the two systems. PCA analyses the whole image, of which hair may contribute quite a large number of pixels and, therefore, overall variance. The graph matching system fits a grid over the face, that concentrates relatively more on inner features and is thus less affected by hair.

Does this mean that graph-matching is a better model of human perception? Not necessarily. First, the pixel-level analysis of the PCA we have used is only one possible instantiation of such a system. We could use PCA based on initially filtered images to give a more physiologically plausible front end (Hancock, Burton and Bruce, 1995). Second, in any case humans also seem to have their memory for unfamiliar faces dominated by pictorial details initially, illustrated here by the significantly better performance in the NN condition. PCA may provide a better account of how humans recognise pictures of previously unfamiliar faces, while graph-matching may provide a closer approximation to the representational processes which eventually allow us to generalise to novel exemplars.

We have here presented a methodology for the comparison of computer-based face processing systems that goes beyond the simple question of how many they identify correctly. We hope to extend our comparisons to include other systems, such as the local feature analysis developed by Penev and Atick (1996) or the related independent components analysis (Bartlett and Sejnowski 1997), as well as to look at the effects of filtration of images prior to the application of PCA (Hancock et al 1995). Since PCA is a just a linear redescription of variables, we might also look at how much it is adding to our analysis beyond the improvements given by separating out the shape.

## Acknowledgements

We are grateful to Kiran Chitta and Derek Carson who ran the experiments with human participants. Norbert Krüger and Michael Pötzsch patiently explained how to use the Bochum face recognition system. Michael Rinne and Michael Pötzsch produced the Gabor plot shown in Figure 4. The paper has been improved following helpful suggestions by reviewers of an earlier version. This research was supported by a SERC grant to Burton, Bruce & Craw (GRH 93828) and an ESRC grant to Bruce and Burton (R000 236688) .

## References

Bartlett, M.S. and Sejnowski, T.J. (1997) Independent component analysis of face images: A representation for face recognition. Proceedings of the 4th Annual

Joint Symposium on Neural Computation, Pasadena, CA, May 1997. La Jolla, CA: Institute for Neural Computation.
Biederman, I. (1998, in press) Neural and psychophysical analysis of object and face recognition, To appear in H.Wechsler, J.P. Phillips, V. Bruce, F. Fogelman Soulie, T. Huang (Eds) Face recognition: From theory to applications. Springer Verlag.
Bienenstock, E. and Doursat, R. (1991) Issues of representation in neural networks. Vision and Vision Research, A. Gorea (Ed), Cambridge MA: Cambridge University Press.
Bruce, V. (1982). Changing faces: visual and non-visual coding processes in face recognition. British Journal of Psychology, 73, 105-116.
Bruce, V. & Young, A. (1986). Understanding face recognition. British Journal of Psychology, 77, 305-327.
Bruce, V., Hanna, E., Dench, N., Healy, P. & Burton, A.M. (1992). The importance of "mass" in line drawings of faces. Applied Cognitive Psychology, 6, 619-628.
Bruce, V. & Langton, S. (1994). The use of pigmentation and shading information in recognizing the sex and identities of faces. Perception, 23, 803-822.
Craw, I & Cameron, P (1991). Parameterising images for recognition and reconstruction. In P. Mowforth (Ed) Proceedings of the British Machine Vision Conference. Berlin: Springer-Verlag
Davies, G.M., Ellis, H.D. & Shepherd, J.W. (1978) Face recognition accuracy as a function of mode of representation. Journal of Applied Psychology, 63, 180-187.
Galper, R.E. and Hochberg, J. (1971) Repetition memory for photographs of faces. American Journal of Psychology, 84, 351-354.
Hancock, P.J.B., Burton, A.M. and Bruce, V. Preprocessing images of faces, correlations with human perceptions of distinctiveness and familiarity. In proceedings of IEE Fifth international conference on image processing and its applications. Edinburgh, July 1995.
Hancock, P.J.B., Burton, A.M. and Bruce, V. Face processing: human perception and principal components analysis. Memory and Cognition 24, 26-40, 1996.
Hill, H. & Bruce, V. (1996). The effects of lighting on the perception of facial surfaces. Journal of Experimental Psychology: Human Perception and Performance, 22, 986-1004.
Johnston, A., Hill, H. & Carman, N. (1992). Recognizing faces: effects of lighting direction, inversion and brightness reversal. Perception, 21, 365-375.
Jones, J. and Palmer, L. (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive-fields in cat striate cortex. Journal of Neurophysiology, 58, 1233-1258.
Klatzky, R.L. & Forrest, F.H. (1984). Recognizing familiar and unfamiliar faces. Memory & Cognition, 12, 60-70.
Lades, M., Vorbrüggen, J.C., Buhmann, J., Lage, J., von der Malsburg, C., Würtz, R.P. and Konen, W. (1993) Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers, 42, 300-311.
von der Malsburg, C. (1985) Nervous structures with dynamical links. Berichte der Bunsengesellschaft für Physikalische Chemie, 89, 703-710.
O'Toole, A.J., Abdi, H., Deffenbacher, K.A. & Bartlett, J.C. (1991). Simulating the "other-race effect" as a problem in perceptual learning. Connection Science:

Journal of Neural Computing, Artificial Intelligence and Cognitive Research, 3, 163-178.
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 American Optical Society A. 10, 405-411
O'Toole, A.J., Deffenbacher, K.A., Valentin, D. & Abdi, H. (1994). Structural aspects of face recognition and the other race effect. Memory & Cognition, 22, 208-224.
Penev, J.S. and Atick, J.J. (1996) Local feaure analysis: a general statistical theory for object representation. Network: computation in neural systems, 7, 477-500.
Pentland, A., Moghaddam, B. and Starber, T. (1994) View-based and modular eigenspaces for face recognition. In Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, 84-91
Phillips, J., Rauss, P.J. and Der, S.Z. (1996) FERET (Face recognition technology) recognition algorithm development and test results, Army Research Lab technical report.
Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3, 71-86.
Vokey, J.R. & Read, J.D. (1992). Familiarity, memorability and the effect of typicality on the recognition of faces. Memory & Cognition, 20, 291-302.
Wiskott, L., Fellous, J.-M., Krüger, N., von der Malsburg, C. (1995). Face Recognition and Gender Determination, in Proceedings of the International Workshop on Automatic Face and Gesture recognition, Zurich.