Face processing: human perception and principal components analysis

Peter J.B. Hancock[*], A. Mike Burton[+], and Vicki Bruce[*]

* Department of Psychology, University of Stirling
+ Department of Psychology, University of Glasgow

Abstract

Principal component analysis (PCA) of face images is here related to subjects' performance on the same images. In two experiments subjects were shown a set of faces and asked to rate them for distinctiveness. They were subsequently shown a superset of faces and asked to identify those which appeared originally. Replicating previous work, we found that hits and false positives (FPs) did not correlate: those faces easy to identify as being "seen" were unrelated to those faces easy to reject as being "unseen". PCA was performed on three data sets: (i) face images with eye-position standardised; (ii) face images morphed to a standard template to remove shape information; (iii) the shape information from faces only. Analyses based upon PCA of shape-free faces gave high predictions of FPs, while shape information itself contributed only to hits. Furthermore, while FPs were generally predictable from components early in the PCA, hits appear to be accounted for by later components. We conclude that shape and "texture" (the image-based information remaining after morphing) may be used separately by the human face processing system, and that PCA of images offers a useful tool for understanding this system.

Address correspondence to Peter Hancock, Department of Psychology, University of Stirling, Scotland, FK9 4LA, UK.

We are grateful to Ian Craw and Nick Costen for providing the images used in these experiments and to Derek Carson who ran Experiments 1 and 2. The manuscript has been improved following comments from Geoffrey Loftus, Mark Reinitz and Peter Dixon. This research is supported by a SERC grant to Burton, Bruce & Craw (GRH 93828).

To appear in Memory and Cognition

Introduction

Psychological research on face recognition has tended to divide into two broad approaches. One approach has been to concentrate on cognitive processes following perception, and to develop information processing models (e.g. Hay & Young, 1982; Ellis, 1986; Bruce & Young, 1986; Burton, Bruce & Johnston, 1990; Young & Bruce, 1991). This approach has been very successful in delineating the stages involved in face recognition, however each of these models has assumed some perceptual processing prior to input. Indeed, some information processing models explicitly require input in the form of componential "face primitives", but remain uncommitted about the nature of these primitives (e.g. Burton, 1994; Valentine, 1991; Farah, O'Reilly & Vecera, 1993).

Other research by psychologists has investigated the perceptual processing of face patterns, demonstrating, for example, how faces seem to be analysed holistically rather than decomposed into discrete local features (e.g. see Bartlett & Searcy, 1993; Tanaka & Farah, 1993; Rhodes, Brake & Atkinson, 1993; Young, Hellawell & Hay, 1987). However, this research tends not to consider the way in which such perceptual processes deliver codes suitable for the task of recognising individual faces. In contrast, a growing body of research by computer scientists and engineers has addressed this question explicitly in the quest for artificial face recognition systems suitable for security and forensic applications . This research has largely progressed without considering the psychological plausibility of the coding schemes employed.

The aim of the work presented in this paper is to examine the psychological plausibility of one such scheme for coding face images for recognition - the Principal Components Analysis (PCA) of face images (Kirby & Sirovich, 1990; Turk & Pentland, 1991). PCA has a number of characteristics which make it attractive as a candidate model of human face image coding, as we elaborate below, and recent work (e.g. O'Toole, Deffenbacher, Valentin & Abdi, 1994) has shown that PCA of a set of face images does a good job of accounting for some aspects of human memory performance with these same images. Our work with PCA builds upon these earlier studies and shows that it is possible to improve the psychological predictive power of a PCA-based model by incorporating a pre-processing stage in which the spatial deviation of each face shape from the average (its "shape") is coded separately.

In this introduction we first consider the evidence that PCA belongs to the right class of image analysis schemes for psychological plausibility, and then consider details of the approach itself. We then consider the way in which a PCA-based system might in principle allow us to implement psychological theories of "face space" and "norm-based" coding. This introduction motivates the new experimental and image analysis work presented in this paper.

Evidence for image-based face coding schemes

There have recently been a number of studies which attempt to understand how the human visual system analyses and stores face images in order to relate image analysis to psychological aspects of face processing. In these studies researchers have evaluated a particular candidate for "face primitives", with respect to human performance data on faces. For example, a number of researchers have examined the potential of simple Euclidean measures such as length of nose, width of mouth and so on (e.g. see Rhodes, 1988). Combinations of such measures taken from a large corpus of faces have been used to derive indices corresponding to human judgements of sex (Bruce, Burton, Dench, Hanna, Healey, Mason, Coombes, Fright & Linney, 1993a; Burton, Bruce & Dench, 1993) and to human judgements of distinctiveness (Bruce, Burton & Dench, 1994). In both these projects, the authors conclude that primitives based on these Euclidean distances alone are probably insufficient to understand internal representations of faces.

A rather different approach, influenced by the 3D model-based approach to visual object recognition (e.g. Biederman, 1987) was taken by Bruce, Coombes and Richards (1993b). Bruce et al (1993b) examined the psychological plausibility of a 3D surface-based coding scheme - where each face was described as a spatial distribution of 3D surface primitives such as peaks, pits, valleys and ridges. While it can be shown that variations in surface descriptions co-vary with psychological dimensions (Bruce et al, 1993a,b), the observation that face recognition is highly error-prone when faces are displayed as surface images devoid of texture or pigmentation (Bruce, Healy, Burton, Doyle, Coombes & Linney, 1991) suggests that the "face primitives" used for recognition of faces cannot be based upon 3D shape descriptions alone.

For a coding scheme to have psychological plausibility it must be able to account for the difficulty that people have with recognising faces shown in certain formats. For example, recognition is extremely difficult when faces are portrayed as line drawings in which major (e.g. mouth) and minor (e.g. wrinkles) face "features" are traced (Davies, Ellis & Shepherd, 1978; Bruce, Hanna, Dench, Healy & Burton, 1992). It is difficult to explain why such drawings are so difficult to recognise if our coding of faces is based upon Euclidean metric measurements, which should be preserved in such drawings. Recognition of line drawings of faces improves dramatically if, as well as "edges", they contain information about areas of relative dark and light from the original image (Bruce et al, 1992). Similarly, face recognition is dramatically impaired if faces are shown in photographic negatives (Bruce et al, 1993a; Hayes, Morrone & Burr, 1986), even though a negative image of a face preserves the spatial layout of the face. Such observations suggest that human facial image coding incorporates information about image intensities themselves, and not just the spatial layout of changes in image intensity. The relative pattern of light and dark within a face conveys important discriminating information about such things as hair and skin colour, and 3D shape from patterns of shading and shadows.

Principal Components Analysis (PCA) is one example of a scheme which codes image intensities and which does not decompose faces into localised "features". O'Toole and her colleagues have performed PCA on facial images and related these to human performance on recognition of own and other race facial images (O'Toole et al, 1994), and on human performance in sex judgements (Abdi, Valentin, Edelman & O'Toole, submitted). Results from these studies have been promising; it appears that PCA may provide a plausible candidate for the notion of "facial primitives".

The PCA approach and shape-free PCA
The basic technique of PCA on images of faces is now well-developed (Kirby & Sirovich, 1990; Turk & Pentland, 1991). A set of facial images is collected and registered (e.g. by normalising the position of the eyes for each face). These images may then be considered as a one-dimensional array of pixel values (grey levels). Correlations are taken between these images, and the coefficients of the principal components (eigenvectors, sometimes known as eigenfaces) are extracted. The coefficients have the same dimension as each of the input images, and may be displayed, see figure 1. As the images are pre-processed to have a zero mean, the eigenfaces code deviations from the mean, and have a rather ghostly appearance. The first eigenface codes the direction of maximum variance in these images. The second codes the direction of maximum variance, after that accounted for by the first has been removed. They are difficult to interpret visually but some features may be observed. Looking at the forehead area of the images in figure 1, the first and third components appear to reflect overall fringe length, while the second and fourth are lopsided, corresponding to individuals with hair on only one side of their forehead. Note that the sign with which the images are displayed is not significant, so we cannot say that the first component makes the whole forehead lighter, while the second makes the left (as we look at it) side darker, only that the first is symmetric, the second not. Note also that the later eigenfaces apparently carry more fine detail. This is consistent with the suggestion by O'Toole et al (1994) that the later components carry information about identity.


Figure 1. The first four, the 20th and the 40th eigenfaces generated from the complete set of 174 full images.

Each of the faces in the corpus generating the components may then be reconstructed by a weighted sum of the eigenvectors. Similarly, new faces may be stored as a weighted sum of eigenfaces. This offers a mechanism for compact storage of face images. A full PCA of a set of 100 faces will generate 99 components (plus the mean pixel values). Because the early components capture most of the variance, it may be possible to produce visually acceptable regenerations of the images from only, say, 50 components. This almost halves the required storage, requiring only the 50 eigenface images, and then 50 component values for each face. New faces may then be coded using the existing 50 eigenfaces, requiring storage of only the 50 component values for each additional face. How well a new faces is regenerated will depend on its match to the original corpus. A face that differs significantly, for instance in race, may well be rather poorly coded. This characteristic of PCA will be exploited in the studies reported here.

The compact principal component coding of a face forms the basis of PCA-based face recognition. The match between a novel face image and an existing database is performed in the reduced space of the coded images. Our interest here is not an attempt at face recognition per se, but an investigation of the psychological plausibility of eigenface-based representation of faces. Some researchers have already noted that particular eigenfaces appear to code particular facial characteristics. For example, O'Toole, Abdi, Deffenbacher & Bartlett (1991) have shown that information about the sex of a face may be present in the early eigenfaces (those with the largest eigenvalues).

While analysis based upon image characteristics may have psychological plausibility, it brings with it a number of problems which arise from the specific characteristics of the images used to create the corpus of faces. It is usually acknowledged that some preprocessing of images is required, for example to normalise intensity values. In the experiments which follow we present data from a different type of preprocessing. Before subjecting face images to PCA, we first eliminate any deviation they have from the average shape of the corpus. This technique was introduced by Craw & Cameron (1991) and is illustrated in Figure 2. One first defines a set of key points which are located for each face. An average value is calculated for each of these points, and a grid is constructed, corresponding to the average face-shape. The same points are used to construct a grid for each facial image. This grid is then morphed into the averaged shape using simple interpolation. The result is an image which has standard shape, and is called a "shape-free face" by Craw & Cameron.

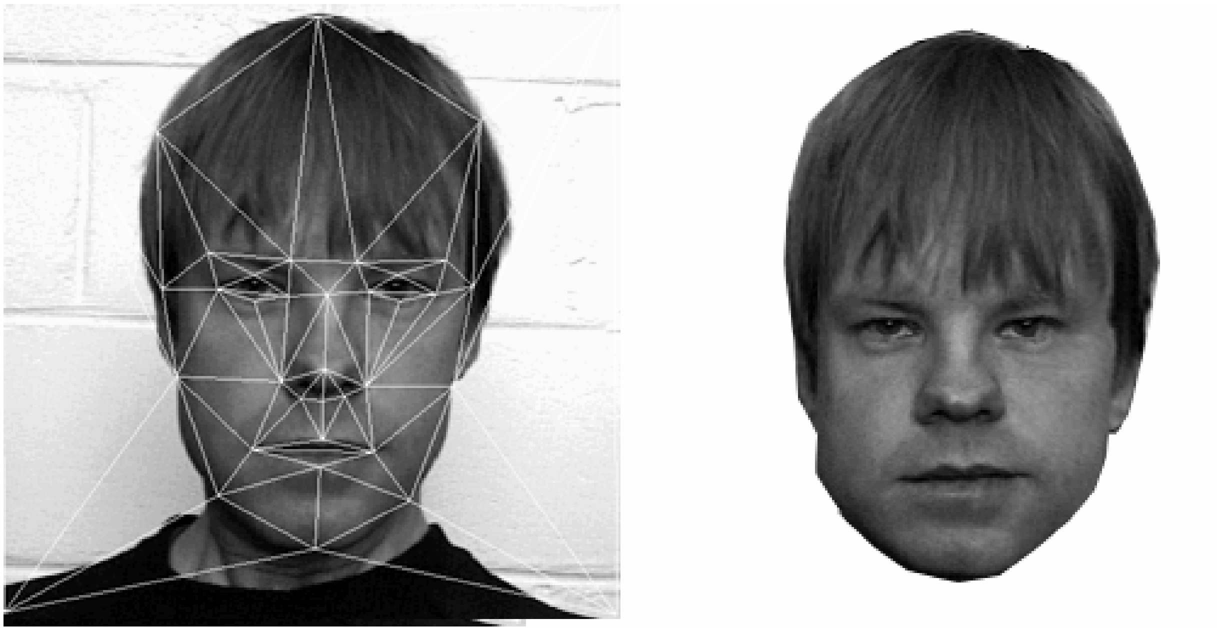

Figure 2. a) The set of control points used to define the positions of features around each face. b) The same face, morphed to the average shape, with background removed.

There are several computational advantages of using shape-free faces for PCA. Previous researchers have subjected whole images to principal components analysis, these images including backgrounds which are sometimes noisy (e.g. Turk & Pentland, 1991). Once all faces have the same shape, it is simple to separate face from background and to analyse only the face. Second, when reconstructing a face the components derived from shape-free faces may be added in linear fashion without changing the overall shape of the face. Combination of components from shaped images leads to blurred edges due to contributions from components derived over a range of face shapes. Finally, in extracting the shape from a face, one can independently examine the contributions of shape - the spatial deviation of each face from the average - and "texture". We use the word "texture" in this paper as a short-hand to cover all the image information which remains in the shape-free face, i.e. colour information and the fine scale features unaffected by the shape averaging. It is important to note that our use of the word "texture" is more restricted than its more conventional usage in psychology. As we will elaborate next, the use of PCA to derive a set of "dimensions" along which faces vary, and the separate analysis of shape and "texture", provides a possible means of implementing psychological theories of "face space" and "norm-based coding" using a coding scheme which is sensitive to low-level image properties.

Face space and norm-based coding.

Much of the recent literature on face recognition has (often implicitly) relied on the notion of "face space", i.e. that there are a number of dimensions along which faces vary, and that a face can be uniquely represented as a point, or vector, in that space. This notion was made explicit by Valentine (1991), who (among other effects) provided an account of facial distinctiveness in these terms. The phenomena associated with distinctiveness in face recognition are well documented. Unfamiliar faces which subjects have rated as being distinctive tend to be better remembered (in a recognition paradigm) than faces which have been rated as typical. Familiar faces which have been rated as distinctive tend to be recognised (as familiar) faster than familiar faces which have been rated as typical (Valentine & Bruce, 1986a). The explanation of these effects in terms of face space proceeds as follows. Faces rated as typical will tend to have common values on dimensions defining face space. This means that typical faces will be clustered together. Distinctive faces, on the other hand, will tend to be relatively isolated in face space: distinctive faces, by definition, have few faces which look similar. Valentine (1991) argues that the relative isolation of distinctive faces makes them easier to recognise than typical faces, as there will be fewer competitor faces in the relevant region of face space.

This discussion of face-space has proceeded without any reference to the actual nature of the dimensions along which faces vary: what are the dimensions of this space? The central aim of this paper is to examine the possibility that the dimensions along which faces vary can be captured in a principal components analysis of images. This is a very different approach to that taken by other workers in this field. For example, discussing the nature of "face space", Valentine (1991) states that "previous work using multidimensional scaling techniques suggest that the principal dimensions needed would represent hair colour and length, face shape and age" (p. 166). In this paper we examine the possibility that faces may be coded on dimensions extracted from PCA on face images, rather than corresponding to these common sense dimensions. Of course, this does not rule out the possibility that dimensions extracted from PCA may themselves have a strong relation with dimensions such as hair-length used in everyday life descriptions of faces.

Among adherents to this view of face space, there is disagreement about the nature of representation. Valentine (1991) has contrasted "norm-based" and "exemplar-based" coding. Norm-based coding refers to the idea that faces are encoded as a vector in face space, with reference to a central norm calculated as the "average" of the known population of faces. This idea has been used to account for caricature effects in face recognition. Rhodes, Brennen & Carey (1987) showed that faces distorted away from a central mean may, in some circumstances, be recognised more accurately (faster) than a veridical image of a face (see also Benson & Perrett, 1991). This has led to the suggestion that faces are coded as deviations from a central tendency of one's known population of faces. In contrast to norm-based coding, some researchers have used the idea of exemplar-based coding (Nosofsky, 1986) to capture the notion that faces are coded in face-space, but without reference to a central norm. Valentine & Endo (1992) have produced some preliminary evidence, based on the other-race effect, that faces may be coded without reference to a central norm.

In the absence of any concrete proposals about the nature of underlying dimensions of face space, it is difficult to separate predictions from the norm-based and exemplar-based views. In this paper we will not address this distinction. However, it appears to us that the use of principal components analysis on shape-free faces does, for the first time, offer the possibility of a computational account of norm-based coding. If there is, indeed, a central norm for our set of known faces, then one way in which faces deviate from this norm is in their different shapes. In the analyses presented in the second half of this paper we shall show that it is possible to examine this type of deviation independently from deviation due to the "texture" of a face (e.g. the variance due to different coloration). By pre-processing facial images to remove variance in shape (and by separately analysing this information) it is possible to measure variability from a central norm due to these different types of information.

In this paper we test the ability of a PCA-based coding scheme operating on whole faces and separately on "shape" and "shape-free" faces, to account for variations in the rated "distinctiveness" of face images, and to account for variations in human memory performance with the same face images. However before describing these investigations we must introduce some complications which arise in the relationship between rated "distinctiveness" and measures of memory.

Distinctiveness

It is most people's intuition that there should be a straightforward mapping between the dimension which subjects respond to when asked to rate face typicality/distinctiveness, and memory performance with faces. Faces which are highly distinctive in appearance (e.g. a face with a long red beard) should - one might think - be highly memorable and rarely give rise to false alarms when presented as distractors. Faces which are very typical in appearance should be less memorable but more likely to give rise to false alarms. In fact this intuition turns out to be wrong. Vokey & Read (1992) discovered that rated "typicality" (cf. distinctiveness) is in fact composed of two orthogonal components, one coding familiarity, and another coding memorability. By decomposing subjects' ratings of faces, they showed that the tendency to rate a face as familiar ("context-free familiarity") dissociates from the tendency to rate the face as being memorable. This finding was replicated by O'Toole et al (1994). These researchers found that for faces from one's own race, ratings to the question "Is the face confusable with someone you know?" dissociate from ratings to the question "Is the face easy to remember?" Bruce, Burton and Dench (1994) revealed a similar dissociation in memory performance rather than in ratings. Rated distinctiveness of faces correlated positively with hit rates to these items, and negatively to rates of false positives (FPs) when the faces served as distractors, but there was a zero correlation between hit rates and FPs to the same items.

Bruce et al's (1994) studies used pictures of faces devoid of hair. In this paper we start by replicating the finding of two orthogonal components of typicality using more natural images of faces shown with hair. We then go on to examine the possibility that PCA decomposition of facial images captures the important dimensions on which faces vary. We relate this decomposition to human performance on the same images. Our particular aim is to establish whether the two independent components of typicality (memorability and familiarity) can be accounted for independently by PCA decomposition. In the course of this exploration, we shall examine whether separate components of a face reflecting its shape and its surface "texture" give rise to different effects in the two components of typicality.

Experiment 1

The aim of this experiment was to gather distinctiveness and memorability ratings on a set of faces, to be used for the subsequent image analysis using PCA.

Materials

174 black and white photographs of young adult males, normalised for inter-ocular distance and eye position were selected from the Aberdeen Frame Face Database (e.g. Shepherd, 1986) The people photographed had no facial hair or spectacles, a

neutral expression, and their clothing was concealed by a dark gown tied at the neck. Photographic subjects were looking directly at the camera, in diffuse lighting.

Subjects
34 volunteer students, male and female, were paid to take part in the experiment.

Method
The image set was divided, at random, into two sets of 87 images, set A and set B. Subjects were asked to rate the faces in one of these subsets for distinctiveness by answering the question: "How easy would it be to spot this person at a train station?". The rating set was preceded by a familiarisation set of 9 faces drawn from the same population but not used in further analysis. This served three purposes: to orientate subjects on the type of faces and likely range of distinctiveness to be used; to acquaint them with the methodology (clicking a response box with a mouse pointer); and to reduce primacy effects in the following recall stage. Responses were made on a scale of 1-10. Subjects were allowed to study each face for as long as they wished. Presentation order was randomised independently for each subject. This is the task used to collect ratings of distinctiveness in several previous studies (e.g. Valentine & Bruce, 1986a, 1986b; Bruce et al 1994).

Following this rating stage, subjects were asked to take part in a separate experiment (on object recognition) lasting about 10 minutes. They were then unexpectedly presented with the complete set of 174 faces, in sequence, and asked for each face "did you see this person before?". Subjects responded on a 10 point scale: 1 = certain I did not see the face before, 10 = certain I did see the face before. Once again, presentation of test faces was randomised independently for each subject.

This technique allows a number of direct measures to be taken for <u>each face</u>. First, a mean distinctiveness rating, derived from the subjects rating faces in a particular set. Second, a measure of "hits" corresponding to the certainty score of subjects who actually saw the face in the learning phase. Third, a measure of "false positives" corresponding to the certainty score of subjects who did not see the face in the learning phase.

Results

	Distinctiveness	Hit score	False positive score	d'
Set A	5.67	6.95	3.58	1.48
Set B	5.8	6.96	3.79	1.25

Table 1: Mean scores for the two sets of faces

Table 1 shows the mean ratings for subjects exposed to each half of the set of faces. The two sets show very similar levels of distinctiveness, "hit" and "false positive" ratings. Table 1 also shows estimates of d' and criterion. These estimates are calculated by taking hit and false positive scores for each face, counting responses of 6 and above as positives and responses of 5 and below as negative. The table shows mean d' for faces which appeared in each of the two sets. Because we are averaging the results of single observations from each subject, the d' values are likely to be underestimated. However, the average d' value of 1.37 is remarkably similar to the value of 1.36 reported by O'Toole et al. using the same calculation for a very different set of faces.

	Distinctiveness - Hit	Distinctiveness - False positive	False positive - Hit	Distinctiveness - d'
Set A	0.55	-0.39	-0.17	0.51
Set B	0.4	-0.42	0.06	0.5
Both	0.49	-0.4	-0.08	0.5

Table 2: Correlation between subject responses (critical value r = 0.21 for sets A & B, 0.15 for combined set)

Table 2 shows the correlation between the average subject responses for each face, broken down by subset, and jointly. These correlations once again demonstrate that while distinctiveness is correlated both with hit and distractor ratings, hit and distractor ratings are themselves uncorrelated. Figure 3 shows scatterplots of the data. This replicates a study using a different set of faces with hair concealed (Bruce et al, 1994), and again provides further support for the dissociation first described by Vokey & Read (1992). We have since replicated this pattern of data in two further studies in our laboratory, one of which used much smaller memory sets (16 targets and 16 distractors) so it does not seem to arise as a consequence of the large memory load. Furthermore, as we describe below, the dissociation does not appear to arise from order effects.

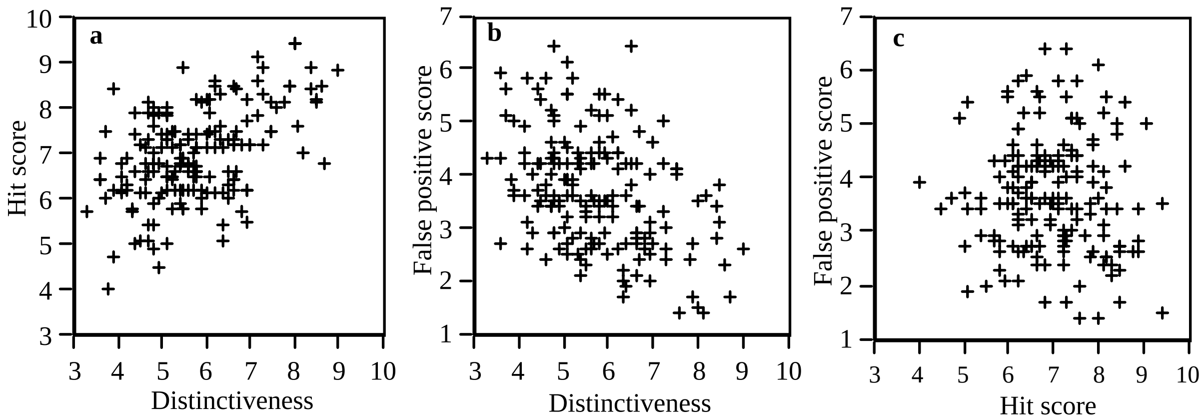

Figure 3. Scatterplots for all 174 faces showing a) positive correlation between distinctiveness and hit score. b) negative correlation between distinctiveness and false positive score. c) Lack of correlation between hit and false positive scores.

Further analysis
In later sections of this paper the images used in Experiment 1 will be analysed using PCA. The intention is to establish whether PCA provides any account of the psychological dimensions revealed here. However, before describing these image analyses, we briefly report some further analyses of the data from this experiment.

Dimensions of subject performance The striking finding in the data presented above is that hit and false positive scores dissociate. This means that faces which are easy to remember if subjects have seen them are not necessarily those faces which are easy to reject if subjects have not seen them. This finding has been treated in different ways by researchers in the past. Bruce, Burton & Dench (1994) tried to account directly for the dimensions corresponding to hit and false positive scores observed in human data. In contrast, Vokey & Read (1992) used factor analysis on their subjects' ratings of typicality, memorability, attractiveness, familiarity and likeablity (this is the standard use of PCA to do data reduction, not PCA on face images) to derive two orthogonal components which they label "memorability" and "context free familiarity".

In the analyses presented below, we have examined both direct and derived measures of human performance. In order to explore the structure of the data presented above, we performed factor analysis on the rating and performance data from these studies (note the contrast with the purely rating data of Vokey & Read, and of O'Toole et al, 1994). From the three scores available from each face (distinctiveness rating, hit scores and false positive scores), we extracted two orthogonal factors. Table 3 shows the correlations between these factors and the subjects' rating and performance scores.

	Distinctiveness	Hit	False Positive	d'	Criterion
Factor 1 (Memorability)	0.93	0.66	-0.6	0.74	-0.06
Factor 2 (Familiarity)	0.08	0.60	0.75	-0.16	-0.9

Table 3: Correlations between two orthogonal factors and the human data (critical value r = 0.15)

The first factor is heavily loaded onto distinctiveness, and also onto d', thereby justifying its label of memorability. As with the study by Vokey & Read, we appear to have extracted a component which codes the ability of subjects to recall having seen a face. The second factor is uncorrelated with both distinctiveness and d', but heavily loaded onto criterion - the subjects' response bias for each face. This appears to capture the dimension labelled context free familiarity by Vokey & Read, i.e. it seems to capture the tendency for subjects to claim that they have seen the face before, irrespective of whether or not it appeared in the learning set.

As there is no very good reason to claim that either the direct subject data (distinctiveness, hits, false positives) or the derived data (memorability and context free familiarity) should act as the standards against which subsequent performance of image analysis should be measured, we examine the relation between PCA on images and both of these types of measure in later sections.

Between subject variability Before attempting to relate subjects' data to the properties of the facial images, it is important to examine the consistency of subjects. There are two reasons why this is necessary. First, we need to know whether there is general agreement about whether a face is distinctive, and whether the same faces are rejected or remembered. A general agreement would indicate (at least the possibility) that the measured dimensions are something to do with the faces themselves, rather than with idiosyncratic subject variables. Second, we need a measure against which to examine our subsequent analysis. Any analysis derived by synthetic means can only be expected to correspond to subject data to the same extent as there is agreement among subjects.

In order to examine the consistency of subjects, we carried out further analysis of the data from distinctiveness, hit and distractor scores. There are two plausible ways to examine consistency. Table 4a shows the mean correlations between every pair of subjects within each group (each mean being comprised of 136 correlation coefficients). Table 4b shows a measure of consistency of sub-totals of subject performance. Each group was split into two subsets (of 8 and 9 subjects), and the mean scores for each subset were correlated. This procedure was repeated 12 times with different random divisions into subsets. Table 4b shows the mean correlations derived from this measure.

		Distinctiveness	Hit	False Positive
a)	Pairwise Set A	0.28	0.09	0.13
	Set B	0.23	0.04	0.17
b)	Grouped Set A	0.79	0.45	0.63
	Set B	0.71	0.28	0.56

Table 4: Inter-subjects correlations, experiment 1. 4a shows the mean of all pairwise correlations. 4b shows the mean of subset correlations produced by random split.

Table 4 shows an initially surprising result. The subjects appear to be relatively consistent when assigning distinctiveness scores. On the other hand, they are much less consistent in which faces they remember from the learning phase. The measures in table 4 show a consistent pattern. The measure of distinctiveness has highest consistency between subjects, followed by the false positive score, with hits showing the lowest consistency across subjects.

Despite this generally stable pattern of consistency, the overall levels are disappointingly low. The demands of this task were quite high on subjects; recall that subjects first rate a set of 87 faces, and subsequently have to rate a superset of 174 faces for whether they have been seen before. It is possible that the very large numbers of faces seen may give rise to inconsistent behaviour. As the order of presentation was randomised independently for each subject, any effect of tiring would affect faces in an inconsistent manner.

We considered two artifactual reasons why subjects should fail to agree on hits. The first was the time spent studying each face during the rating phase. It seems plausible that faces looked at longer for some reason would be better recalled. However, analysis showed only one subject for which there was a significant correlation between observation time and hit score. Viewing time can therefore be eliminated as an explanation.

The second possible reason for the low consistency of hit scores is the randomised order of presentation. Although the initial display of nine faces in the familiarisation phase should reduce possible primacy effects, it still seemed possible that recency, or residual primacy effects might be affecting recall. A plot of order of presentation during the rating phase against hit score showed no such list-end effects, and we concluded that this was not a significant source of inter-subject difference. However, a plot of hit score against presentation order during recall showed a significant negative correlation ($r=-0.38$, $p<0.05$), as did a plot of false positive scores ($r=-0.32$, $p< 0.05$). A possible explanation for this is that subjects were aware that half the test faces had been presented for rating. Since many of the faces are quite similar, subjects might tend to respond positively at the start of the test, but increasingly negatively as their subjective ration of hits became depleted. Whatever the reason, it would contribute to the observed lack of correlation between subjects. We therefore replicated Experiment 1, but using fewer items and consistent ordering.

Experiment 2.

Method

A subset of 80 of the previous faces were used: the top 10 and the bottom 10 faces rated for distinctiveness from Experiment 1, and 15 each from the four intermediate ranges (4-5, 5-6, 6-7, 7-8). The ends of the range are thus over-represented, relative to the population for Experiment 1. 12 volunteer student subjects were recruited. None had taken part in Experiment 1. The procedure was exactly as in Experiment 1, except that the faces were presented in the same order on each occasion.

Results

As with Experiment 1, we present the data on consistency of subjects in two ways (see Table 4). Table 5a shows the mean correlations for each pair of subjects (15 pairs for each set). Table 5b shows the grouped averages: each group of 6 was split in half in each of the 10 possible ways to give average ratings values for each of the scores.

		Distinctiveness	Hit	False Positive
a) Pairwise	Set A	0.35	0.25	0.30
	Set B	0.42	0.14	0.21
b) Grouped	Set A	0.68	0.44	0.59
	Set B	0.68	0.32	0.44

Table 5: Inter-subjects correlations, experiment 2. 5a shows the mean of all pairwise correlations. 5b shows the mean of subset correlations produced by splitting groups in half.

As expected, the pairwise results are higher than in Table 4a. Removing the order effects and reducing the set size has increased subject consistency. The grouped results are lower than those of table 4b, because of the smaller group sizes. However, the pattern of results is exactly the same as with Experiment 1: subjects are most consistent in their distinctiveness ratings, followed by false positive scores, and least consistent on the hit rates. The same pattern of correlation is observed between the three sets of data for each face: significantly positive between distinctiveness and hit score ($r=0.27$, $p< 0.05$), negative between distinctiveness and false positive ($r=-0.36$, $p< 0.05$) and not significant between hit and false positive scores ($r=-0.21$, $p>0.05$).

Discussion: Experiments 1 and 2

In these experiments we showed that hit rate and false positive scores were uncorrelated in subject data. This puzzling effect seems well-grounded, as it has been demonstrated in ratings data as well as performance data, and we have replicated our performance data in several different studies, including one with only 16 faces in the rating set. In the next section we attempt to account for this effect.

In Experiments 1 and 2 we have also shown that there is a robust pattern in the consistency of subjects. Subjects are in relatively high agreement about which faces are distinctive, behave less consistently in false positives, and are least consistent in their hits. This leads us to make the following tentative proposal. It is possible that subjects make distinctiveness ratings on the basis of two dimensions, and that these are reflected in the hit and false positive scores. The larger consistency of false positive scores may reflect a general property of these faces - perhaps their similarity to the population of faces as a whole. The hit score on the other hand, may reflect idiosyncratic knowledge of faces by subjects. It is possible that certain subjects remember particular faces because they are similar to someone known to the subject. This may explain the inconsistency of these ratings.

This suggestion is clearly speculative, and we shall return to a discussion of these issues at the end of the paper. We now turn to an analysis of the images used in these experiments.

Analysis of images

In the sections above we have presented data from subjects who have been shown sets of facial images. In this section we shall present analyses of these images themselves. The aim is to establish whether characteristics of these images can be found which can account for patterns in the human data. As described in the introduction, we are particularly interested in the possibility that PCA of facial images can reveal aspects of faces which predict psychological effects. This issue has been addressed by O'Toole et al (1994) in respect of the other race effect. These researchers found that the advantage for recognition of faces from one's own race can be captured in PCA analysis of images, assuming a larger exposure to one particular race. In the present paper, we aim to examine the relation between PCA and the psychological data on distinctiveness. In particular, we aim to explore the possibility that PCA can account for the separate effects of hits and false positives, as described above. In addition to extending the range of psychological data addressed by PCA, we shall also examine the possibility that the separate analysis of shape and "texture" may inform this analysis.

The aim of PCA is typically to reduce the dimensional size of the input set. Given the procedure of ordering components, most of the variance is captured by the early factors. In the studies which follow, we shall usually consider only the first 20 eigenfaces, as our observations are that the first 20 dimensions are sufficient to capture most of the variance in the input set. However, we shall analyse this more explicitly in a later section.

Having performed PCA on images, there are two types of information which we will use to relate the images to psychological data. First, we can analyse the PCA outputs for a particular face: what are the weights allocated to it for each derived eigenface? We refer to this set of values as the face's <u>spectrum</u>. Second, we can ask how good is the coding of a particular face. If, for example, we were to reconstruct a face from its spectrum of (say) the first 20 outputs, how well would the face be reconstructed? We refer to the measure derived in this way as the <u>reconstruction error</u> of a face.

In order to conceptualise this more clearly, consider the psychological dimension of distinctiveness. PCA captures a particular notion of distinctiveness: that of being a long way from the mean value. Specifically, it identifies those axes with maximal variance. If, in capturing variations from the mean in either the image grey levels or the measured shape of the face, PCA captures something of what our subjects regard as distinctive, then faces rated as distinctive should tend to have large component values. In other words, we might expect that distinctive faces are allocated values which are distant from the mean on some of the components. If we now consider reconstruction errors, one might predict that distinctive faces are coded less well by the early eigenfaces, precisely because the early components capture variations common to many faces, which distinctive faces, by definition, do not share so much. Distinctive faces should therefore have high reconstruction errors. We shall investigate both these possibilities in the following sections, as well as considering the relation between PCA and the other psychologically-derived measures.

Finally, there is an added complication in considering PCA of images. The discussion above assumes that the same set of images is used to generate PC coefficients and for subsequent testing. However, it is equally plausible to generate principal components with one set of images, and to code a new set on the components generated. In this case we refer to the outputs as the reflection of the second set through the components of the first. This procedure may be seen as more similar to the human experiment, where experience derived from previously seen faces is brought to bear on a novel set. It can be used to generate data for either the spectrum or reconstruction error analyses described above, and in the following we shall examine both same-set and reflection data.

PCA analysis offers a possible explanation of the dissociation between well-remembered and well-rejected faces. It might be expected that faces that are badly coded by the system (human or computer) would be poorly remembered. Such a face would be regarded as distinctive, in the sense of having a high reconstruction error. It should be easy to reject such a face, on the grounds that, by definition, it is outside the range of familiar faces. These, then, would be the faces that are distinctive and well-rejected but poorly remembered. Conversely, a high PC spectrum output from a face that is distinctive, but within the space coded by the coefficients, might correlate with those that are better remembered due to particular similarity to individual faces known by the subject. Unfortunately this is difficult to test experimentally, since the population of known faces will differ between subjects. For a distinctive face in the test set to be well-coded by the PCs would require another similar distinctive face to be present in the generation set, which, given the modest size of the sets, seems unlikely.

Materials: Image processing For computer processing, the 256x256 images shown to the subjects were reduced to 64x64 with 256 levels of grey. In the analyses which follow, PCA is performed on both untransformed and shape-free images. To generate the shape-free images, a shape map for each face was defined manually, specifying the x,y coordinates of 35 locations of features such as eyes and nose, and the periphery of chin and hair (see Figure 1). These coordinates were used to produce shape-free face images, by "morphing" to the average shape, using bi-linear interpolation. The background of both shaped and shape-free images was removed by setting the pixels outside of the area bounded by the shape map to zero.

1. Multiple regression

We first attempt to account for the human data using multiple linear regression. The approach adopted is to use PC outputs (i.e. spectrum values) for each face to predict the human data. Taking distinctiveness as an example, we might expect there to be large correlations (multiple-R) between the absolute spectrum values and distinctiveness. This is because, as explained in the previous section, we might predict that faces rated as distinctive will have large (discrepant) values on some or all of the components.

Intuitively, large absolute values of PC outputs might contribute to a face's distinctiveness, since distinctive may lie both sides of the mean. However, in this study we are also attempting to predict other human data. We shall use PC outputs to predict the data collected on hit and false positive scores, and also to predict the derived dimensions labelled "memorability" and "context free familiarity" (see above). In contrast to distinctiveness, there is no simple intuitive relation between these scores and PC outputs. We therefore used both absolute and raw (signed) PC outputs as predictor variables in the following study.

a) Whole face set In the first Multiple Regression study, we used the first 20 components derived from a PCA on the entire 174 image set. The decision to use 20 components was somewhat arbitrary, based on the observation from initial tests that there was little sign of consistent correlations between any higher components and the subject rating data. Both raw and absolute values were entered as predictor variables. Variables were entered by stepwise addition, with the criterion that to enter a variable must increase the multiple-R significantly ($F > 3.84$).

In order to study the relative contributions of shape and "texture" to these correlations, we repeated this procedure four times, using predictor variables from PCA on four different sources. These were:

1. The original images, adjusted to bring their eyes to the same coordinates.
2. The shape-free images, morphed to fix the coordinates of 35 locations.
3. The shape vectors - the set of 35 {x,y} pairs for the morphing coordinates.
4. The first 13 components from the shape free images and the first 7 from the shape vectors.

Data set 1 gives a measure of this technique for untransformed facial images, while data sets 2 and 3 analyse face "texture" and face shape separately. Data set 4 represents an attempt to re-combine shape and "texture" after separate PCA on these two sources. If these two aspects of a face are analysed separately, then each must undergo the PCA before being brought together. This ratio of "texture" to shape components (approximately 2:1) was chosen by inspection of the correlations given by individual component outputs, which were mostly low for shape components higher than 7 (see also Table 8). Note that the shape vector has only 70 entries per image. Extracting 20 components from this set accounts for much more of the total variance than 20 components of the 4096 dimensional image data. We used the same number of components to facilitate comparison between the resultant multiple-R values.

	Distinctiveness	Hit	False Positive	Memorability	Familiarity
1. Full	0.51	0.33	0.36	0.51	0.34
2. Shape Free	0.40	0.28	0.50	0.40	0.40
3. Shape vector	0.42	0.19	0.25	0.44	0.22
4. Shape free + shape	0.49	0.36	0.44	0.43	0.40

Table 6: Multiple regression MR values for PC outputs (spectrum) predicting human data, using whole set of faces.

The multiple-R values between the component outputs and the human data for these four cases are shown in Table 6. With 40 variables, it is to be expected that some apparently significant correlations will arise by chance. An estimate of this chance level of multiple-R was obtained by randomising the order of the human data and rerunning the multiple regression. This was repeated 100 times, giving an average multiple-R of 0.28, which we take to be chance performance.

There are several points to note from this table. Distinctiveness is predicted at reasonably high levels by all the different types of data. As might be expected, removing the shape from the images reduces the level of prediction achieved (multiple-R = 0.51 vs 0.40), and the shape-alone and shape-free images achieve roughly equivalent levels of prediction (0.40 vs 0.42). Hit scores are predicted poorly (at chance) by the shape-free and shape-alone data, and better by the full (untransformed) data, and by the combination of shape and shape-free data. Perhaps the most surprising results come from predictions of false positives. It seems that removing the shape from a face makes a substantial improvement in the correlation with false positive scores (multiple-R = 0.36 vs 0.50, F test, $p< 0.01$). Further, the shape vector alone does not predict FP rates at above chance levels. Finally, the derived components memorability and familiarity behave similarly to distinctiveness and false positive respectively. Using some components from both shape vector and the shape free images appears to give the best of both, with correlations for all the subject data being near their best.

We are unable to claim too much from these results, due to the dangers inherent in multiple regression. These are highlighted by the results for hit score, one of which is considerably below the value expected by chance. Small variations in the data may make the difference between a variable entering the equation or not, with consequent effects on the reported multiple-R. Although differences such as those between 0.51 for distinctiveness and 0.33 for hit score with full faces or 0.40 for distinctiveness and shape free are formally significant (F test, $p< 0.01$), the evident noise makes such comparisons unconvincing. We therefore adopted the following method.

b) Random segmentation of the face set The set of faces was split randomly in half. One half was used to generate PC coefficients. These coefficients were then used to analyse the other half. The outputs obtained were used to perform multiple regression on the corresponding human data. This process was repeated 100 times, to obtain average multiple-R values. It is not usually safe to average correlation coefficients, because of their non-normal distribution; furthermore, the samples are not independent, being drawn from the same complete set. However, we have no reason to suppose that the distribution of correlations for the various human data will be different. Averaging should therefore not affect the rank ordering of the correlations obtained, and we shall concentrate on this ranking below. Averaging serves to smooth out the effects of random correlations, and produces a clearer pattern of results, shown in Table 7.

	Distinctiveness	Hit	False Positive	Memorability	Familiarity
1. Full	0.51	0.42	0.42	0.48	0.38
2. Shape Free	0.48	0.40	0.49	0.44	0.44
3. Shape vector	0.48	0.37	0.30	0.42	0.32
4. Shape free + shape	0.52	0.43	0.48	0.49	0.42

Table 7: Averaged multiple regression MR values for PC outputs (spectrum) predicting human data. Data averaged over 100 random half splits of the face set.

Estimates of the chance correlation were obtained as before by randomising the order of the human data and rerunning the multiple regression, for each of the 100 segmentations of the data set. They came out very consistently, with none differing significantly from an overall average of 0.336. This number is larger than the value of 0.28 obtained above, because we are only using half the faces on each test. For these randomised controls, ANOVA, comparing 100 runs from each of four preprocessing methods over five subject variables, shows no effect of variable (distinctiveness, hit, etc), $F(4,396) < 1$, no effect of preprocessing (shape, shape-free, etc), $F(3,297) < 1$, and no interaction, $F(12,1188)=1.24$, $p > 0.2$. We therefore assume that the expected chance correlations are the same for all the conditions shown in Table 7 and that the results may therefore be compared directly with each other. All of the multiple-R values shown are significantly above the random value of 0.336 (t-test, $p<0.01$) except those from the shape vector to false positive score and context-free familiarity. Differences between the multiple-R values were tested using the Mann-Whitney U test on the 100 samples for each condition, at $p < 0.01$. To give a feel for the consistency of the results, noting the problems of averaging correlation coefficients, the standard errors on these data are all approximately 0.01.

In summary, we:
1. Split the face set in half at random.
2. Extracted principal components from one half.
3. Reflected the remaining faces through these components.
4. Performed multiple regression between the raw and absolute values of the first 20 component outputs (13 image and 7 shape for the combined type) with each of the 5 subject measurements.
5. Randomised the order of the subject measurements and repeated the multiple regression to obtain an estimate of the chance correlations.
6. Repeated steps 1-5 100 times for each of the four pre-processing types and averaged the results within pre-processing type and subject measurement pairs.

The pattern of results confirms that suggested by Table 6:

1. With the full faces, the results for hit and false positive scores are the same, that for distinctiveness significantly higher.
2. Moving to shape free faces causes a significant drop for memorability, a small downward trend for distinctiveness and hit score, but a significant increase for false positive score and familiarity.
3. The values for shape vector alone are significantly worse than for full-face for all variables. False positive and familiarity are at the random values. Hit score is above chance, but barely.
4. Adding the shape vector PCs to those from shape free images does best of all. The multiple R for distinctiveness is significantly better than that from the shape-free images and indistinguishable from that from the shaped images, while that for false positive is significantly better than the result from shaped faces, and indistinguishable from the shape free performance. The three values for hit do not differ significantly.

So far as we are aware, this is the first demonstration of the possible psychological relevance of shape averaging. The process increases the ability of PCA to extract information that leads people to say yes when they have not seen a face before, which affects both false positive and familiarity scores. Conversely, there appears to be no information available from the shape vector about false positives. So far as we can account for human false positive scores, the information comes from fine detail in the image, which we refer to as the "texture". We shall return to general discussion of these issues at the end of the paper. However, we now examine the contribution of individual components to the psychological predictions described here.

2. Individual component correlations.
The multiple regression results suggest that different aspects of the images carry information about memorability and familiarity. Further insight into the nature of the features underlying these two dimensions might be obtained by examining the loading of the individual principal components. In this section we examine whether particular components carry information specific to particular psychological dimensions.

There are a number of ways that this might be done. The first is to generate components from the complete set of faces, echoing the first procedure for multiple regression above. We may then look at the individual correlations between each component and the subject ratings for each face. Since this can only be done once, there are uncertainties about the replicability of the results.

A second approach is to split the face set randomly as before and produce average correlations for each component. In addition to the usual problems of averaging correlation coefficients, there is an additional problem of potential inconsistencies in the order of principal components. With the set split in half, we have 87 data points in the 4096 dimensional pixel space: clearly a very sparse sampling. Given the relative homogeneity of our images, this may not be too problematic, but we can nonetheless expect the derived components to vary between runs. For instance, the information contained in component 7 on one occasion might appear as component 8 in another, or be redistributed among two or more different components. Visual inspection of the eigenfaces from different runs suggests that this starts happening as early as component 3 or 4. Any correlations between individual components and the subject ratings will therefore be unstable. Averaging the results will lead to correlations being spread over a number of neighbouring components.

This problem of uncertain distribution of variance may also affect the first method (using all the images to generate components). It could be that there is useful information about distinctiveness, say, that usually occurs around component 10, but on this occasion is distributed amongst several other components. These will each show a small, apparently non-significant correlation, where a slight change would lead to a significant correlation for component 10. Averaging would allow such effects to show through.

A third approach is to look at the components actually used in the multiple regression equation. This has the advantage of identifying whether the various components are accounting for different aspects of the subject ratings. The first four components might all correlate with distinctiveness, but all be capturing the same part of the variance. The multiple regression equation should include only the most significant of these, giving a clearer indication of the most important components. As before, we may look either at the regression equation for the whole set of faces, with consequent doubts about repeatability, or use the random sampling technique. The same problem about inconsistent components will apply: over a number of runs we would expect to find neighbouring clusters of components, at most one or two of which appear in any particular regression equation.

Results Table 8 shows the usage of components in the 100 multiple regression equations generated for the previous section. Almost all the components occurred at least once, as would be expected because of chance correlations. During the 100 control runs with randomised subject ratings, each component occurred on average 4.67 times. The binomial distribution then requires more than 12 occurrences in 100 trials for $p<0.001$ (such a low p value being used because we have 40 variables). Table 8 reports the components that occurred more than 12 times.

Type	Distinctiveness	Hit	False positive
Full	3, 4, 7, 8, 9, 10, 12, A1, A2	4, 6, 9, 10, A1, A2	1, 2
Shape free	2, 3, A1, A15, A17	2, 7, 17, A2	1, 2, 6, 7, A5
Shape vector	4, 6, 19, A1, A7, A9, A18	2, 4, 6, A7	A7
Combined: shape	1, 4, 6, A1, A6, A7	4, 7, A1, A7	A7
image	2, A1	2	1, 2, 6, 7, A5
	Memorability	Familiarity	
Full	3, 4, 8, 9, 10, 11, 19, A1, A2	1	
Shape free	2, A1, A15	1, 6, 7, 19, A5	
Shape vector	4, 6, A1, A7, A18	4, 6	
Combined: shape	1, 4, 6, A1, A7	-	
image	2, A1	1, 6	

Table 8. Usage of components in multiple regression equations, from same 100 runs as Table 7. Components listed occurred more often than expected by binomial distribution ($p<0.001$). A1 refers to absolute value of component 1, etc.

The pattern of component usage confirms the multiple regression results: false positive and familiarity showing an increase in components used when going from full image to shape free, while the other three show a decrease. The relative usage of information from the shape vector and the image is shown clearly by the combined component results, with false positive and familiarity loading heavily onto the image components, the other three onto shape components. Although our familiarity measure is derived about equally from the hit and false positive scores (Table 3), it behaves more like the latter. Only the use of components 4 and 6 from the shape vector echoes the loading of hit score.

Within each pre-processing category, there is little in common between false positive and the other two direct subject ratings. Thus false positive uses the first two components from the full image set, while distinctiveness and hit use several in the range 3-12. A striking result not indicated by Table 8 is the frequency with which false positive used the first component: 73 of the 100 runs for the full images and 82 for the shape free images. The other "frequent" occurrences are typically in the range 20-40. One reason for this may simply be that it is the first component, and therefore relatively stable. As indicated above, it may be that there is a similar amount of information about distinctiveness that occurs somewhere in the range 7-10 for the full images, inconsistently because of the variability on the principal components. Further analysis of the data supports this suggestion, showing that at least one of these components occurs in 78 of the 100 equations for distinctiveness and that never more than two of them occur together.

3. Reconstruction errors

We now turn to another measure of PCA performance: reconstruction error. This gives us a measure of how well a particular face is coded, when included in the whole data set. The purpose of this analysis is to explore the possibility that distinctive faces are coded less well by PCA than are typical faces.

O'Toole et al (1994) reported reconstruction error using a normalised cosine error, where 1 means perfect reconstruction and smaller numbers are worse. A possible alternative is simple Euclidean distance between the input and reconstructed images (i.e. the length of the vector between the two points in 4096 dimensional image space representing the face and its reconstruction). If the image vectors are normalised to unit length before computing the distance, this measure differs from O'Toole's only in the cosine non-linearity. Tests showed much larger correlations if the vectors were not normalised.

Although our stated aim is to examine whether distinctiveness can be captured as a correlate of "goodness of coding" in PCA, the goodness of coding is itself dependent on a number of parameters. In particular, the more components that are extracted from the set, the better the general coding. In this study we calculated the reconstruction error for different numbers of extracted components. These measures are then correlated with the psychological data.

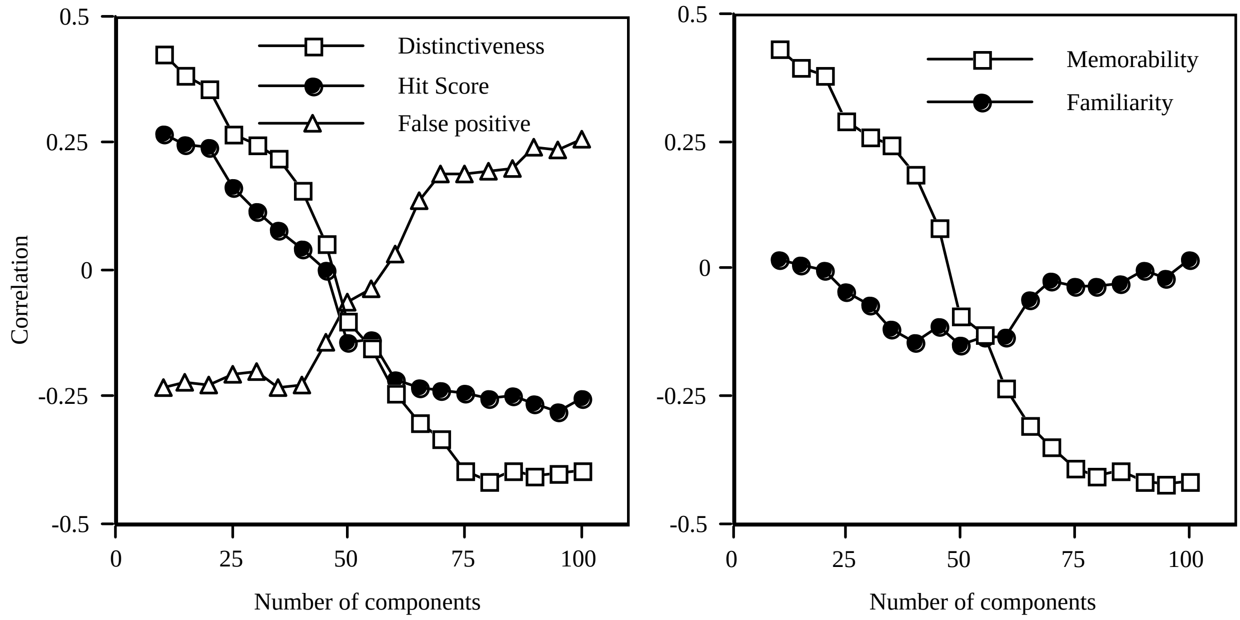

Figure 4. Correlation between the reconstruction error and the subject ratings for each face as the number of components used for the reconstruction is varied.

Figure 4 shows the correlation between the unnormalised Euclidean reconstruction error and each of the three direct and two derived subject data sets as the number of components used for the reconstruction is varied. The complete set of 174 full (i.e. not shape-free) images was used both to generate the components and to test reconstruction. Note that we are considering far more than the 20 components used for the multiple regression studies: while higher components showed little direct correlation with the subject data, they all have a (gradually diminishing) effect on the reconstruction error.

To understand Figure 4, consider the first (leftmost) points. If only a small number of components are used to reconstruct a face, then the correlation between reconstruction error and distinctiveness is high. This means that faces rated as highly distinctive are reconstructed badly. However, the correlation between reconstruction error and false positive score is negative: i.e. faces that are easy to reject, and so have a low false positive score, are reconstructed badly. Typical faces, that are poorly remembered and rejected, are relatively well coded, with low error.

The striking result shown in these graphs is the reversal of the sign of most of the correlations as the number of components is increased. This is most easily understood by considering distinctiveness. Each additional component will accommodate as much of the remaining variance as possible. Initially this is best done by coding features common to many images. Since lack of shared features is one definition of distinctiveness, these early components code average faces better than unusual ones. Distinctive faces therefore have a high reconstruction error. As the number of components is increased, there comes a point when there is little variance common to several images left to be accommodated. Now the best strategy for reducing variance is to cover those distinctive faces that were poorly coded by the early components. The correlations reverse, so that above about 50 components, distinctive faces are better coded than those that are more average. If the analysis were continued beyond 100 components towards the number of face images used, the correlation would fall back to zero along with the reconstruction error.

Distinctiveness and (derived) memorability behave almost identically, with hit score qualitatively similar, but at lower correlations. False positive score behaves almost like a mirror image, but crosses zero at closer to 60 components. The derived factor familiarity here behaves as a combination of hit and false positive. Correlation is close to zero, with only a mild negative excursion in the middle range, where hit score has dropped and false positive is still negative.

General Discussion

The studies described above were performed to analyse the relationship between human face processing and structural properties of face images. Like previous work in this field (O'Toole et al, 1994, 1991; Abdi et al, submitted) we have attempted to capture psychological effects in terms of the statistical properties of images, as revealed by PCA. We now summarise the data, and offer some conclusions.

The psychological data on hits and false positives do not correlate. This means that faces which are easy to recognise as having been seen, are not necessarily those which are easy to reject when they have not been seen. Some of the data described above suggest that we may be able to separate properties of the face which give rise to these two dimensions.

The studies of reconstruction error show that when one uses only the early components, faces which are badly coded have a high hit rate; whereas faces which are well-coded have a high false positive rate. In other words, faces which are discrepant on these components are easy to recognise as having been seen, while faces which are not discrepant on these dimensions are those most likely to be falsely identified as having been seen. When one uses a large number of components this pattern reverses. This seems to indicate that it is the early components which give rise to false positives, whereas later components give rise to hit rate. This tentative conclusion is supported by the studies examining the individual components (Table 8). These show that, for full face images, it is the very early components which are most commonly used to predict false positive, whereas the components which predict hits tend to be drawn from later in the spectrum.

These suggestions are consistent with an intuitive notion of what is captured by the early and late components derived from PCA. The early components code very general information, extracting information common to all faces in the set. In general we might say that these components define the range of face-like patterns (of pixel intensities). However, later components begin to pick up individual variation, as is shown by the reversal of effects in Figure 4. It is to be expected that hit and false positive scores, which are uncorrelated, would therefore load onto different, inherently orthogonal principal components. It is intuitively sensible that it should be the false positive score that loads onto the earliest components.

Further support for this view can be found in the separate analysis of shape and shape-free faces. Data from the multiple regression studies show that the shape-free faces capture the false positive data best of all (Tables 6 and 7). In other words, it is variation in what we have called "texture" which gives rise to false positives; variation in shape seems to make no contribution to the chances that a face will be falsely recognised.

Once again, examination of the data from individual components seems to support this. Table 8 (fourth line) shows the contribution of different components to the data from combined shape and shape-free information. Components from the shape-only information appear to load specifically onto hit scores, whereas components from the image (shape-free) information load onto false positive scores.

It appears quite clear from the data presented here that false positives arise as a consequence of a face's similarity to the general population. Furthermore, the measure of similarity used in this account does not include information about face shape, but rather information about coloration or "texture". The interpretation of data from hits (and correspondingly from distinctiveness overall) is less easy to explain. It appears that dimensions giving rise to hits do include some information about the shape of a face. This lends some support to the notion of norm-based coding as discussed in the introduction. Deviations away from an average shape appear to predict the accuracy with which subjects identify a face as having been seen. However, shape does not account for the whole effect. Table 7 shows that the best predictor of hits (and distinctiveness) arises from separate analysis of shape information and "texture" information, subsequently brought together.

In the discussion of Experiments 1 and 2, we suggested that hit scores may be partly determined by subjects' idiosyncratic knowledge of people. Perhaps subjects score hits partly because a particular face reminds them of an acquaintance. This is consistent with the low levels of subject agreement on hit scores, and with the general finding that the later (more detailed) components tend to load on hits. We are now conducting experiments to test this hypothesis further.

In conclusion, it appears that we have isolated some of the separate information which gives rise to psychological properties of face perception. In particular, we have concentrated on subjects' hit and false positive scores in remembering faces. We have not done this by breaking down images of faces into everyday components like noses, chins etc. Rather, we have extracted statistical properties of the images. Furthermore, we have shown that a decomposition of these images into separate shape and "texture" information makes it easier to account for some of the psychological data, suggesting that this distinction may also be made by the human system.

Although this paper reports some success in relating human data to statistical properties of images, we have not provided evidence that the PCA performed here is the best way to capture this data. The PCA was performed directly on the image pixels, for simplicity and because this is the approach taken by other workers. However, it is clear that the human visual system performs various types of filtration relatively early in its processing. It is quite possible that a more complete account would rely on PCA on filtered images and we have begun to investigate the effects of such pre-processing (Hancock, Burton & Bruce, 1995).

Finally, the analyses presented here may be taken to support the general notion that image-based statistics provide some insight into human face processing. This contrasts with accounts based on local distance measures or surface-based measures as described in the introduction. However, PCA is just one of a broad range of image-based statistical techniques available (e.g. see Brunelli & Poggio, 1993; Wurtz, Vorbruggen & von der Malsburg, 1990, for alternative techniques). Further research is required in order to establish whether the results described here are tied to the particular PCA analysis chosen, or are a general consequence of any image-based technique.

References

Abdi, H., Valentin, D., Edelman, B. & O'Toole, A.J. (submitted). More about the difference between men and women: Evidence from linear neural network and principal component approach.

Bartlett, J.C. & Searcy, J. (1993). Inversion and configuration of faces. Cognitive Psychology, 25, 281-316.

Benson, P.J. & Perret, D.I. (1991). Perception and recognition of photographic quality facial caricatures: implications for the recognition of natural images. European Journal of Cognitive Psychology, 3,105-135.

Biederman, I. (1987). Recognition by components: A theory of human image understanding. Psychological Review, 94, 115-147.

Bruce, V. (1994). Stability from variation - the case of face recognition. Quarterly Journal of Experimental Psychology, 47, 5-28.

Bruce, V., Burton, A.M. & Dench, N. (1994). What's distinctive about a distinctive face? Quarterly Journal of Experimental Psychology, 47, 119-141.

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

Bruce, V., Coombes, A & Richards, R. (1993b). Describing the shapes of faces using surface primitives. Image and Vision Computing, 11, 353-363.

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., Healy, P., Burton, M., Doyle, T., Coombes, A. & Linney, A. (1991). Recognising facial surfaces. Perception, 20, 755-769.

Bruce, V. & Young, A. (1986). Understanding face recognition. British Journal of Psychology, 77, 305-327.

Brunelli, R. & Poggio, T. (1993). Face recognition - features versus templates. IEEE
Transactions on Pattern Analysis and Machine Intelligence, 15, 1042-1052.

Burton, A.M. (1994). Learning new faces in an interactive activation and competition model. Visual Cognition, 1

Burton, A.M., Bruce, V. & Dench, N. (1993). What's the difference between men and women? Evidence from facial measurement. Perception, 22,153-176.

Burton, A.M., Bruce, V. & Johnston, R.A. (1990). Understanding face recognition with an interactive activation model. British Journal of Psychology, 81, 361-380.

Craw, I. & Cameron, P. (1991). Parameterising images for recognition and reconstruction. In P. Mowforth (Ed.) Proceedings of the British Machine Vision Conference, 1991. Berlin: Springer Verlag.

Davies, G.M., Ellis, H.D. & Shepherd, J.W. (1978) Face recognition accuracy as a function of of mode of representation. Journal of Applied Psychology, 63, 180-187.

Ellis, H.D. (1986) Processes underlying face recognition. In R. Bruyer (Ed), The Neuropsychology of Face Perception and Facial Expression. New Jersey: Lawrence Erlbaum Associates Inc.

Farah, M.J., O'Reilly, R.C. & Vecera, S.P. (1993). Dissociated overt and covert recognition as an emergent property of a lesioned neural network. Psychological Review, 100, 571-588.

Hancock, P.J.B., Burton, A.M. and Bruce, V. (1995) Preprocessing images of faces: correlations with human perceptions of distinctiveness and familarity. To appear in proceedings of IEE fifth international conference on image processing and its application, Edinburgh, July 1995.

Hay, D.C. & Young, A.W. (1982) The human face. In A.W. Ellis (Ed.) Normality and pathology in cognitive functions. London: Academic Press.

Hayes, T., Morrone, M.C. & Burr, D.C. (1986). Recognition of positive and negative bandpass-filtered images. Perception, 15, 595-602.

Kirby, M & Sirovich, L. (1990). Applications of the Karhunen-Loeve procedure for the characterisation of human faces. IEEE: Transactions on Pattern Analysis and Machine Intelligence, 12,103-108.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 139-57.

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., Deffenbacher, K.A., Valentin, D. & Abdi, H. (1994). Structural aspects of face recognition and the other race effect. Memory & Cognition, 22, 208-224.

Rhodes, G. (1988). Looking at faces. First-order and second-order features as determinants of facial appearance. Perception, 17, 43-63.

Rhodes, G., Brake, S. & Atkinson, A.P. (1993). What's lost in inverted faces? Cognition, 47, 25-57.
Rhodes, G., Brennen, S. & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. Cognitive Psychology, 19, 473-497.
Shepherd, J. (1986). An interactive computer system for retrieving faces. In H. Ellis, M. Jeeves, F. Newcombe & A. Young (Eds.) Aspects of face processing. Dordrecht: Martinus Nijhoff.
Tanaka, J.W. & Farah, M. (1993). Parts and wholes in face recognition. Quarterly Journal Of Experimental Psychology 46(A), 225-245.
Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3, 71-86.
Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. Quarterly Journal of Experimental Psychology, 43A, 161-204.
Valentine, T. & Bruce, V. (1986a). Recognising familiar faces: The role of distinctiveness and familiarity. Canadian Journal of Psychology, 40, 300-305.
Valentine, T. & Bruce, V. (1986b). The effects of distinctiveness in recognising and classifying faces. Perception, 15, 525-536.
Valentine, T. & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. Quarterly Journal of Experimental Psychology, 44A, 671-703.
Vokey, J.R. & Read, J.D. (1992). Familiarity, Memorability, and the effect of typicality on the recognition of faces. Memory and Cognition, 20, 291-302.
Wurtz, R.P., Vorbruggen, J.C & von der Malsburg, C. (1990). A transputer system for the recognition of human faces by labeled graph matching. In Eckmiller, R., Hartmann, G. & Hauske, G. (Eds). Parallel Processing In Neural Systems And Computers. pp 37-41.
Young, A.W. & Bruce, V. (1991). Perceptual categories and the computation of grandmother. European Journal of Cognitive Psychology, 3, 5-49.
Young, A.W., Hellawell, D. & Hay, D.C. (1987). Configural information in face perception. Perception, 16, 747-759.