Cognitive Psychology. July 2005

Robust representations for face recognition: the power of averages

A. Mike Burton (1), Rob Jenkins (1), Peter J.B. Hancock (2) & David White (1).

1. University of Glasgow, UK
2. University of Stirling, UK

Address correspondence to:

Mike Burton
Department of Psychology
University of Glasgow
Glasgow G12 8QQ
UK

mike@psy.gla.ac.uk

Acknowledgement: This work was supported by an ESRC Grant (R000238357) to Mike Burton and Vicki Bruce, an ESRC Grant (R000230437) to Mike Burton & Rob Jenkins, and a British Academy Postdoctoral Fellowship to Rob Jenkins. We are grateful to Geoffrey Loftus, Alice O'Toole & Tom Busey for helpful comments on a previous version.

Abstract

We are able to recognise familiar faces easily across large variations in image quality, though our ability to match unfamiliar faces is strikingly poor. Here we ask how the representation of a face changes as we become familiar with it. We use a simple image-averaging technique to derive abstract representations of known faces. Using Principal Components Analysis, we show that computational systems based on these averages consistently outperform systems based on collections of instances. Furthermore, the quality of the average improves as more images are used to derive it. These simulations are carried out with famous faces, over which we had no control of superficial image characteristics. We then present data from three experiments demonstrating that image averaging can also improve recognition by human observers. Finally, we describe how PCA on image averages appears to preserve identity-specific face information, while eliminating non-diagnostic pictorial information. We therefore suggest that this is a good candidate for a robust face representation.

Robust representations for face recognition: the power of averages

Introduction

Human face recognition is often assumed to be generally accurate, but in recent years it has become clear that performance is in fact radically different for familiar and unfamiliar faces. To illustrate this key contrast, consider the 'line-up' displays in Figure 1, reproduced from Bruce, Henderson, Greenwood, Hancock, Burton, & Miller (1999). Bruce et al.'s line-up task represents a best-case scenario for identifying images captured on CCTV. For each display, observers are asked to decide whether or not the target face at the top (a still from a high quality video recording) is present in the line-up below (high quality studio photographs), and if it is, to point out the match. This seemingly straightforward task turns out to be surprisingly difficult when the faces are unfamiliar (see Figure 1). Bruce et al. (1999) reported error rates of 30% for those arrays in which a target is present (with subjects claiming no match on roughly 20% of occasions, and choosing the wrong face on roughly 10%). For arrays in which the target was absent, subjects incorrectly chose a match on roughly 30% of occasions, despite being fully informed that targets would be absent in half the arrays.

These results are particularly striking, since the viewing conditions were optimized in a way that could never be met in a real video security system. All images were of good quality, in very similar poses, and under good quality lighting conditions. Furthermore, all images were taken on the same day, eliminating minor differences in hairstyle, weight and health that cause faces to change in appearance, even when the person is not trying to disguise their identity. Bruce et al conclude that the use of CCTV security systems for matching identity is likely to be limited by human perception, just as much as it is limited by technical issues of image quality. This conclusion is consistent with earlier research by Kemp, Towell & Pike (1997) showing that retail assistants find it very difficult to match shoppers to their photo IDs, when the shoppers are unknown to them.

By contrast, the same matching task becomes trivial when familiar faces are used. In fact, this basic contrast runs much deeper. Even though unfamiliar face recognition can often be defeated by this superficial image change (i.e., a change in source camera only), familiar face recognition can survive all manner of profound changes in image, including changes produced by speech, emotional expression, facial hair, make-up, aging, diverse lighting conditions and different characteristics of the camera. Some of these changes are captured in Figure 3a, which shows ten different pictures of the same person. Considering the huge variation among these images, it is difficult to see what they could possibly share that signals the same identity. Nevertheless, familiar face recognition is highly accurate and robust, even when the quality of the image is severely degraded (e.g., Harmon, 1973; Sergent, 1986; Burton, Wilson, Cowan, & Bruce, 1999).

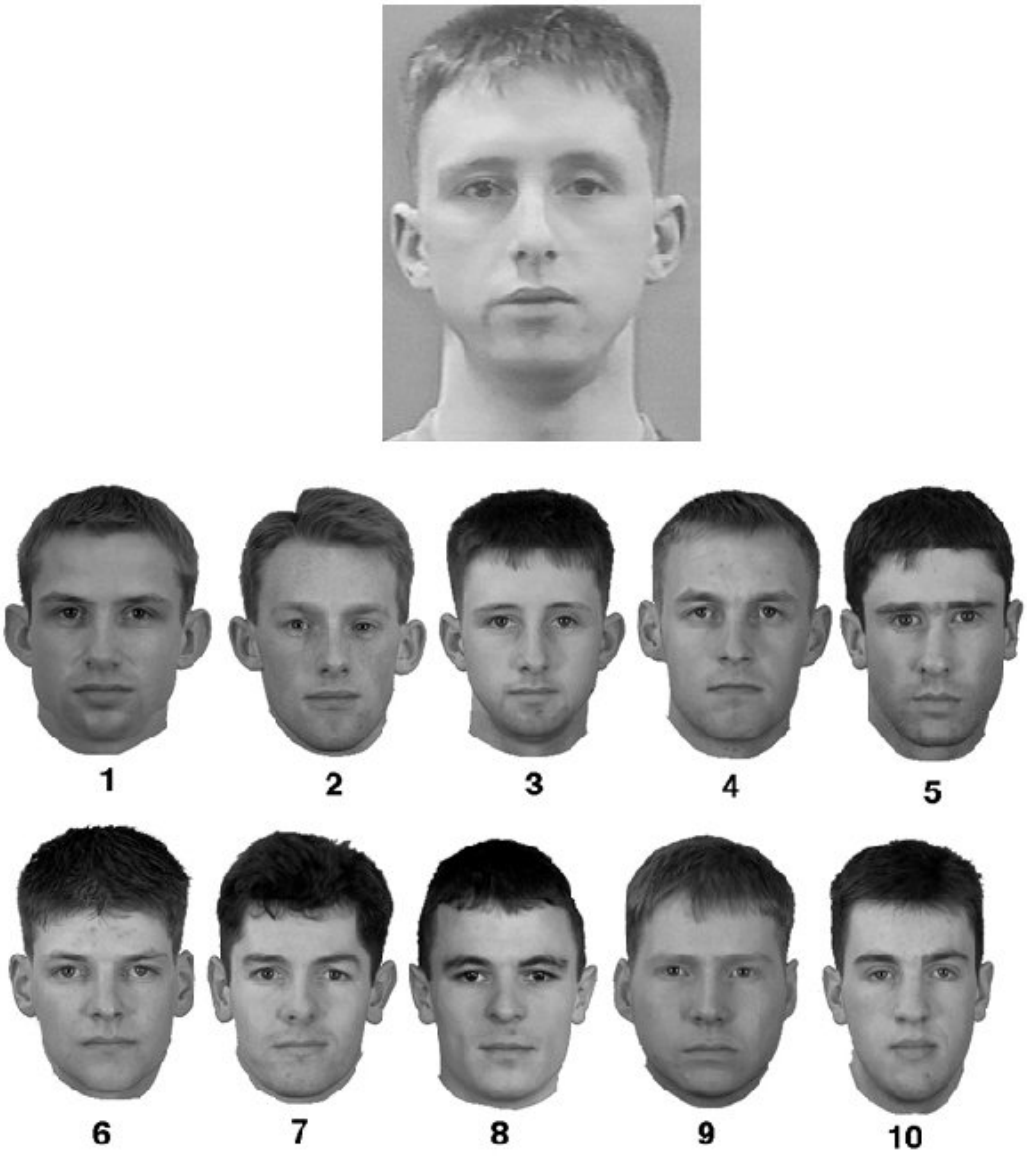

Figure 1a: The person shown at the top may or may not be one of the ten below. Subjects' task is to decide if he is present, and if so, which is he. (Reproduced from Bruce et al, 1999). *Answer given in Appendix.*

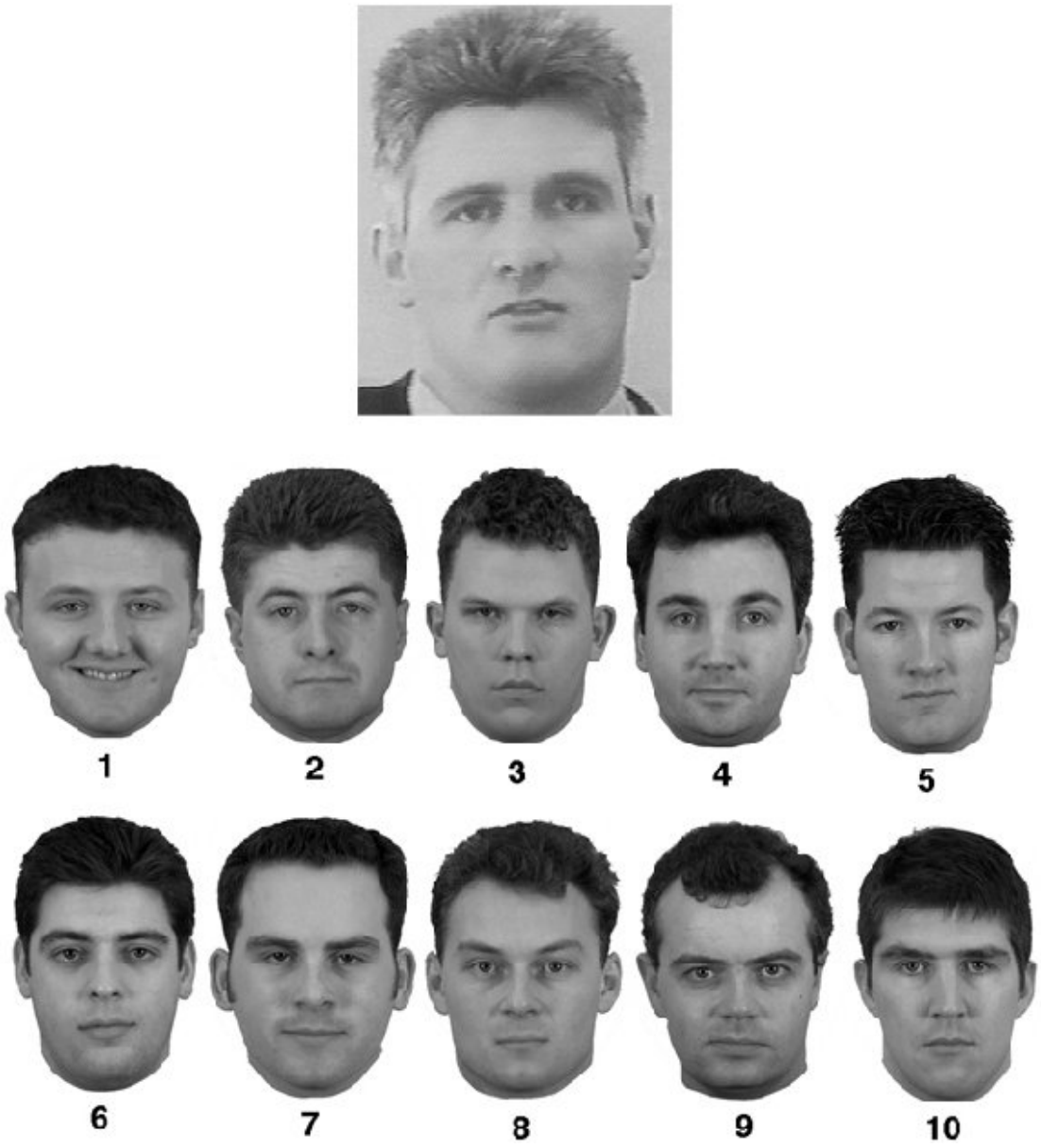

Figure 1b: The person shown at the top may or may not be one of the ten below. Subjects' task is to decide if he is present, and if so, which is he. (Reproduced from Bruce et al, 1999). *Answer given in Appendix.*

This contrast between familiar and unfamiliar face recognition is particularly intriguing given that it must also apply to individual faces over time; every familiar face was unfamiliar when first encountered, and so has presumably undergone a shift from being poorly recognized then to being well recognized now. Here we ask what could drive this shift. To date, the common approach has been to posit a gradual shift towards a more efficient matching strategy over the course of familiarization. For example, it is thought

that the internal features of a face come to dominate recognition, as the person becomes more familiar. So, for unfamiliar faces, matches appear to be based on overall face shape, and hair, whereas for familiar faces, matching seems to rely on eyes, noses and mouths (e.g., Ellis, Shepherd & Davies, 1979; Young, Hay, McWeeny, Flude, & Ellis, 1985; O'Donnell & Bruce, 2001; Bonner, Burton & Bruce, 2003). In this paper we develop an alternative proposal that does not involve an explicit shift in strategy, but focuses instead on exposure-driven refinement of the stored representations against which incoming images are matched.

Exposure is clearly an important factor in strengthening familiarity, as the faces that are most familiar to us are the ones that we have seen the most. But what might increased exposure provide that could lead to better recognition? To attempt to address this question here, we develop the notion put forward by Bruce (1994) of "stability from variation", i.e. that the very variable nature of the stimuli (e.g. Fig 3a) allows the perceiver to distil a powerful representation which incorporates those aspects of the stimulus which are pertinent to the task at hand, while discarding the non-diagnostic variability inherent in any particular set of instances.

Consider two broad approaches to visual representation, one based on storage of individual images, and the other based on storage of a single abstract representation, distilled over many images. A system which stores all encounters with a particular face will improve with increased exposure because it will accumulate more possible matches: the more images one stores of Tony Blair, the more likely it is that an incoming image of Tony Blair will find a good match. In an abstractive system, recognition improves because each new instance refines the quality of the representation, and the canonical representation of a face comes to incorporate, somehow, that which is constant across all the many variations of the face ("stability from variation", Bruce, 1994).

Many psychological models of familiar person recognition include the notion of an abstract representation of faces. Bruce and Young's (1986) influential framework incorporates putative Face Recognition Units (FRUs) which respond to *any* recognizable view of a known person. Such units were intended as analogous to logogens (Morton, 1969), and were present in many precursors of the Bruce & Young model (e.g. Hay & Young, 1982; Ellis, 1986), as well as descendents of it (Brédart, Valentine, Calder & Gassi, 1995; Burton, Bruce & Johnston, 1990; Burton, Bruce & Hancock, 1999; Hanley, 1995, Young & Bruce, 1991). These units have been recruited in explanations of a very wide range of phenomena, for example patterns of priming (Ellis, Young & Flude, 1990; Ellis, Flude, Young & Burton, 1996; Schweinberger, 1996; Young, Hellawell & de Haan, 1988), cross modal person recognition (Hanley & Turner, 2000; Schweinberger, Herholz & Stief, 1997) and certain characteristics of prosopagnosia (Burton, Young, Bruce, Johnston & Ellis, 1991; de Haan, Young & Newcombe, 1987; Young & Burton, 1999). However, despite the theoretical utility of this construct, all the papers cited above remain silent about how it might actually be implemented. How might it be possible to build a representation which becomes active on presentation of any recognizable view of a person? For many researchers, and particularly for vision scientists and engineers

wishing to build useful face recognition systems, this question represents the entire problem of face recognition.

Figure 2: Average images of 50 celebrities. Each image is constructed from 20 different photographs (see text for details of procedure). Names of people depicted are given in the Appendix.

In this paper, we offer one way in which a simple abstractionist system could be implemented for face recognition. The representations we develop are based on simple "averages" of face images. Figure 2 shows images of 50 celebrities, which have been formed from 20 different photographs of each person. The procedure for generating these images is described later, but the important point in this preview is to note that the original photographs from which they were formed are very highly variable (as in Figure 3a). We will show how an architecture based on "average" images performs well in an

artificial face-recognition system, and present some evidence that human observers find these abstract representations particularly easy to process.

Throughout, we will contrast abstract representations such as those shown in figure 2, with the constituent images which were used to build them. It is not, of course, our intention to assert that face recognition *must* be abstractionist. In particular, since the representation we offer has some characteristics of a prototype, we are keen to avoid any claims that prototype models of face recognition are inherently superior to exemplar systems. We doubt that such an assertion is ever possible, and it is certainly not from the data we present. However, we will illustrate that one particular way of implementing a prototype system offers a promising approach to the problem, which has various attractive properties for understanding a range of phenomena.

Automatic face recognition

Automatic face recognition is a topic which currently attracts a great deal of attention. However, it is a difficult problem to solve across a realistic variation in images. In the DARPA-sponsored FERET evaluation of face recognition systems (Phillips, Moon, Rizvi & Rauss, 2000), several algorithms performed well when matching two images of a face, taken in the same sitting, with the same camera, but varied expression. For example, recognition rates of 95% are reported for analyses based on Principal Components Analysis (Moghaddam, Nastar & Pentland, 1997; and see below) and on wavelet-based systems (Wiskott, Fellous, Kruger & von der Malsburg, 1997). However, performance was much poorer for images taken on the same day, but with a different camera (80% in the best case, and only 60% in the second; Phillips et al, 1997). Across all systems tested, none scored higher than 60% when matching images taken a year apart. In a more recent test of modern commercial systems (FERET FRVT2002; Philips, Grother, Michales, Blackburn, Tabassi & Bone, 2002), the best available systems scored only 73% on a recognition test using a real-world database of images, even though these were consistent in quality, and taken in known lighting conditions. Although results from studies with consistent illumination and capture conditions are often promising, generalization to realistic levels of image variation has not been reported. In a recent authoritative survey of available automatic systems, Zhao, Chellappa, Phillips & Rosenfield (2003) write "recognition of face images acquired in an outdoor environment with changes in illumination and/or pose remains a largely unsolved problem …. Current systems are still far away from the capability of the human perception system" (ibid., p. 399).

In the work presented below, we have deliberately chosen to study the difficult problem of face recognition across naturally varying images. The stimuli we have used are images of famous people, gathered from the internet (i.e. those celebrities represented in Figure 2). We have no control over the lighting of the original images, nor of other superficial characteristics such as the contrast, perspective, resolution or focal length of the cameras used to take them. A sample, for a single individual, is shown in Figure 3a. Observers in our experiments (such as those reported later) have little difficulty in identifying any of these individual images as being Tony Blair. Nevertheless, it is

difficult when seeing them all together, to imagine what it is that each of these images has in common to allow easy recognition.

The approach we have taken is to apply Principal Components Analysis to this problem. PCA of images has become a popular technique in understanding face processing, both for engineering, and psychological applications. Originally conceived for use in face *recognition* (Kirby & Sirovich, 1990; Turk & Pentland, 1991; Valentin, Abdi & O'Toole, 1994; Burton, Bruce & Hancock, 1999), it has also been used to model face similarity effects (Hancock, Bruce & Burton, 1996), the "other race effect" (O'Toole, Deffenbecher, Valentin & Abdi, 1994; Furl, Phillips & O'Toole, 2002) and analysis of facial expression (Calder et al, 2001; Cottrell, Branson & Calder, 2002; Dailey et al, 2002). The basic methodology is as follows. A training set of images is subjected to PCA, generating a relatively small number of eigenvectors ("eigenfaces" in this literature). The original images are then re-coded in the space of the eigenfaces, giving each image a unique set of coefficients, which act as its signature. Finally, new test images are projected onto the same eigenfaces, and the resultant coefficients are compared to those of each face in the training set, with a hit occurring when the closest match is with the correct identity.

One limitation of the standard PCA approach is that there is often only a single image stored for each identity known to the system. Furthermore, early reports in the literature generally used images from the same source to serve as target and test faces (e.g. same lighting conditions, same camera etc.). This is important, because superficial image characteristics tend to dominate the match, and if these are varied, the system can easily become insensitive to matches of person identity. In the studies reported below, we have used many images of each known person, against which to match an incoming (previously "unseen") image. We have built systems based (i) on exemplars, in which eigenfaces are derived from several individual instances of each face, and (ii) on averages, in which eigenfaces are derived from a simple image mean of each of the instances of a face.

Image averaging is possible, because prior to PCA, we morph all faces to a standard shape, as illustrated later in figure 8. This is performed in a graphics program by over-laying an image of a face with a grid. The points in the grid are positioned over key points (e.g. corners of the mouth, of the eyes etc) for the particular image under study. The face is then deformed (morphed) to a standard shape, which will be used for all faces in the study. In this way, the same part of each image will contain the mouth, the eyes, and so forth. The resultant images are called "shape-free" in the literature. This technique is due to Craw (1995; Craw & Cameron, 1991), and has been shown to improve PCA considerably (Burton, Miller, Bruce, Hancock, & Henderson, 2001). Similar manipulations which allow separate treatment of the shape and image intensity ("texture") of faces, have been developed for a variety of image-processing techniques (e.g., Beymer, 1995; Vetter & Troje, 1995) and this separate treatment has become a common practice. Examples of the shape-free versions of raw face images are shown in Figure 3b, and an average of these is shown in Figure 3c. The technique of averaging together shape-free images of the same person as a way of producing their face prototype

was first introduced by Benson & Perrett (1993). Here we show that such averages can provide an efficient device for robust face recognition.

Figure 3: Ten images of Tony Blair. Figure 3a (top block) shows original images. Figure 3b (second block) shows the results of morphing each of these images to a standard shape. Figure 3c (bottom) shows the image-average of these shape-standardized images.

Study 1: PCA performance using instances and averages

Images: Ten photographic images of 50 celebrities' faces were gathered from the internet. Each picture showed a roughly full-face front-view, and was from a different source (i.e. we did not use images which had apparently been taken in the same photographic session). The resulting pool of images was thus highly variable in terms of superficial photographic characteristics, and captured a range of facial expressions (see, for example, Figure 3a). The use of celebrities is convenient for two reasons: first it is possible to find many different images of each person, and second we planned to use these for recognition by human subjects in later studies. The identities used are shown in Figure 2 (and identified in the Appendix).

Each of the images was rendered in gray-scale and morphed to a common shape using an in-house program based on bi-linear interpolation (see e.g., Gonzalez & Woods, 2002). Key points in the morphing grid were set manually, using a graphics program to align a standard grid to a set of facial points (eye corners, face outline, etc). Images were then subject to automatic histogram equalization.

Method

In separate simulations, PCA was performed on 1, 3, 6 or 9 images of each person. In instance-based systems, these images were coded separately, while in average-based systems, a single image average was computed from the same pictures. In each case, one set of images was reserved for use as a test set, so that recognition rates for novel images of the 50 faces could be assessed as a function of previous exposure to different images of the faces

Instance based systems

Instance-based systems were built by performing PCA on 1, 3, 6 or 9 images of each face, i.e. on 50, 150, 300 or 450 images in total. In each case, 50 eigenfaces were generated, and all training images were projected onto these. Test set images (50, in all cases) were then projected onto the same eigenfaces. Nearest neighbour matches were generated using a Mahalanobis metric, in which all dimensions are normalized prior to computing distances (Craw, 1985, see discussion). The system is regarded as having made a correct identification if the test image most closely matches one of the target images of the same person. A second measure of performance was also taken, based on summed similarity (Nosofsky, 1988, 1991, see below). Summed Mahalanobis distances were calculated between the test face and all instances of target individuals. The system is regarded as having made a correct identification if the test face has the smallest summed distance to target images of the same person.

For each of the 1-, 3-, and 6-exemplar versions, three different simulations were run, using different learning sets. Single exemplar versions were constructed using images 1, 4 and 7. For the three 3-exemplar versions, we used subsets 1-3, 4-6 and 7-9, while for the three 6-exemplar versions we used the complements of these. For the 9 exemplar

version, a single simulation was run using images 1-9. Image 10 was used as a test image throughout.

Image-average systems

Image-averages were formed by taking the arithmetic means (at each pixel) of 3, 6 or 9 images. Image average sets were constructed from the same images as in each of the exemplar sets above (i.e. there were three 3-image average simulations, three 6-image average simulations, and a single 9-image average simulation). 50 eigenfaces were generated from each set of averages. Nearest neighbour matches were carried out in the same way as for exemplar systems, with a hit being recorded when the test image most closely matched the average image of the correct person. In this case (one target image per person), the summed similarity method is equivalent to nearest neighbour, and so only a single set of matching data was generated for average-based systems.

Results and Discussion

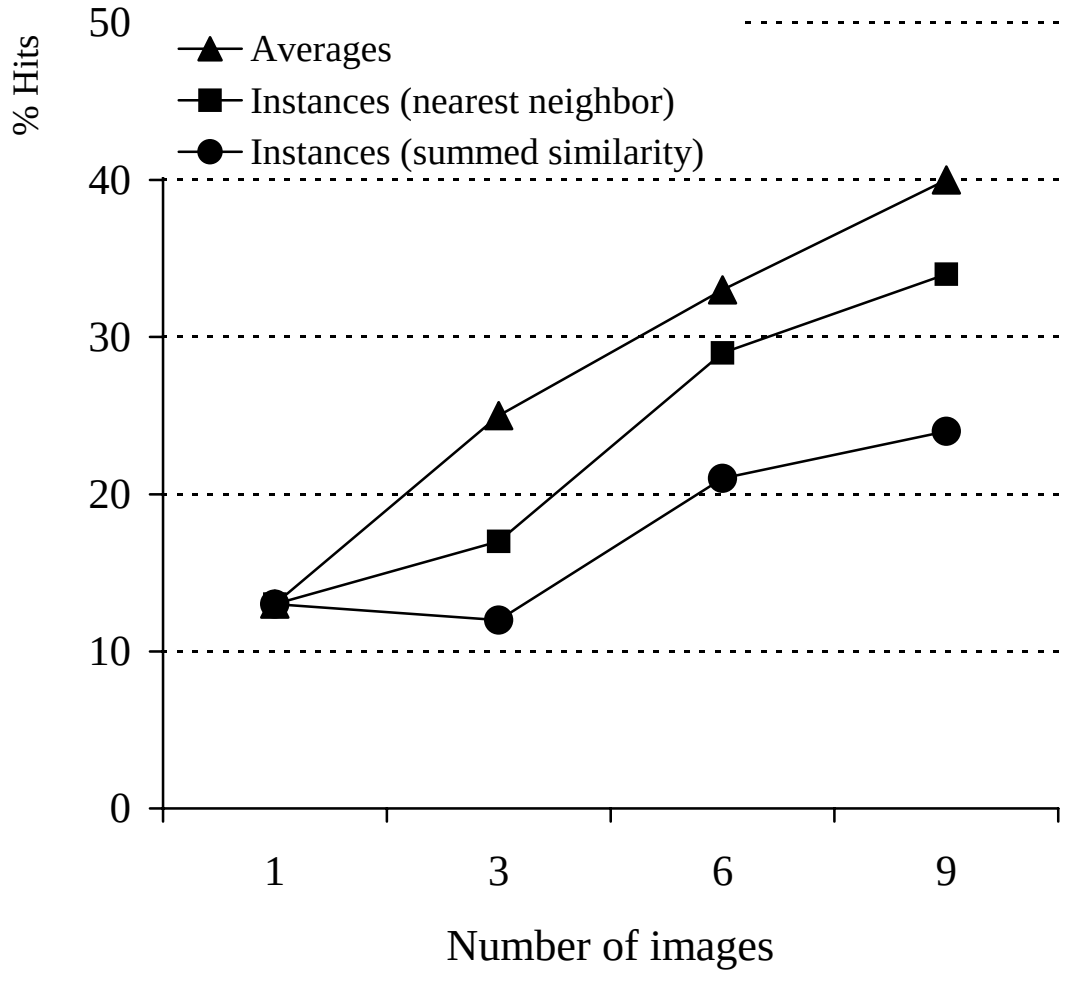

Figure 4: Mean hit rates (%) for systems derived using different numbers of images, for both instance-based and average-based simulations.

Figure 4 shows hit rates for the different systems. Several points are worthy of note. First, the simplest system, using a single training image, performs relatively poorly. PCA systems are rarely reported in which illumination and capture conditions vary widely from training to test sets, which is presumably why the technique has not typically been used with famous faces. Here we have not only variable, but highly variable images contributing to the analysis. Chance hit rate is 2% here, and so performance of 13%, given the huge variation in superficial image characteristics, is possibly better than one might expect. Performance is partly due to the level of standardization of images. The shape-free morphing, plus histogram equalization, brings the images more closely into alignment than the originals. The use of a Mahalanobis distance match is also very important. Under this technique, all dimensions (i.e. principal components) are standardized to have the same variance, prior to Euclidean matching. The technique has been used commonly in PCA research, and has been shown significantly to improve performance, especially when the image source is not held constant across instances (Yambor, Draper & Beveridge, 2002; Burton et al, 2001). The net result is that early components, which capture the largest variance, no longer dominate the match. This is important for images with highly variable superficial characteristics, since these tend to be captured in the early components, but are not diagnostic of identity. [FOOTNOTE 1]

The second point of note is that the image-average systems consistently outperform the instance-based versions, even though the same test images were used across all systems. In every case, it pays to average the training images together rather than to store them separately. Even though instance-based systems provide more target images against which test faces can be matched, performance is better with a single average. Equally important, it seems that the average itself improves as more images contribute to it. So, averaging across nine images per person (40%) is better than averaging across six images (33%), which is again better than using a 3-image average (25%).

These results appear to show that a system based on prototype abstraction (cf. Posner & Keele, 1968), out-performs a system based on storage of instances. This is a complex domain of real images, in which the variance is not under the control of the experimenter, and yet the results are systematic in favouring one technique over the other. Of course, this is a very complex issue in cognition, and many authors have demonstrated that systems based on exemplars can behave in prototype-like ways. The focus of this paper is not to try to distinguish between prototype and exemplar models in general. However, we should note that the prototype advantage demonstrated here is not a trivial consequence of the similarity metric used. Proponents of exemplar models point out that nearest-neighbour matches are particularly susceptible to noise in individual exemplars, and propose the use of summed-similarity metrics instead (Nosofsky, 1988, 1991). In fact, in this study, summed similarity over instances performed consistently worst of all.

This advantage for averages is perhaps surprising, since the target representation is not even a real image of the person it depicts. It lacks some of the characteristics of a real image, taking on a rather soft-focus quality (a point also noted by Benson & Perrett, 1993). It seems then, that a successful match does not rely on fine surface characteristics such as wrinkles, or details of complexion. This may turn out to be an important

component of the prototype advantage. Although images of people certainly do contain a lot of information about superficial fine-scale aspects of the face, if these are not diagnostic of identity, a match without them is likely to improve performance. This notion is, in fact, consistent with research on spatial scale in face recognition, which suggests that identity tends to be carried at low spatial scales (Bachmann, 1991; Harmon & Julesz, 1973; though see Schyns & Oliva, 1997, for an argument that extraction of information from different spatial scales is more flexible when the task is to identify a specific image of a face).

We now appear to have a promising representational technique for matching these highly unconstrained images of faces. We have established that a system based on simple averages provides a reasonable level of performance under the conditions tested here. However, although the performance of this PCA system is surprisingly good compared to chance, it still leaves plenty of room for improvement. In the next study, we set out to establish whether averages built on larger numbers of images would perform any better. We also ask whether it is possible to observe an advantage for averages built of more images within the context of a mixed-level memory, i.e. a situation in which the system knows some identities very well, and others less well.

Study 2: PCA performance as a function of level of familiarity

Most automatic face recognition systems aim to optimize recognition performance on all known faces. However, the human case is clearly more diverse: we know some faces very well indeed, but others much less well. Furthermore, the level of our familiarity with a face is known to predict certain perceptual tasks: simply, the more familiar we are with a face, the more fluent is our processing of it (e.g., Clutterbuck & Johnston, 2002). In the following simulation, we attempt to capture this by building a system based on averages, but in which some averages are constructed from a large number of images, and others are built from fewer images

Method

For this study, a larger pool of 1000 images was used. This pool comprised 20 images of each of the 50 celebrities used in Study 1, all gathered from the internet, and taken in capture conditions over which we had no control. All images were morphed to the same standard shape, and pre-processed as in Study 1. Identity averages were now generated by taking the average across 3, 6, 9 or 19 images of each person, with the 20th image (selected at random) used as a test image for all versions of the system.

Following a similar procedure to Study 1, we performed PCA on 50 images, in which each image corresponded to an identity of one of the known individuals. For ten of these people, the image was a specific instance. The remaining images comprised ten averages constructed from each of 3, 6, 9, and 19 images. The resulting set therefore comprised 50 known identities, with “familiarity” varying from ten people encountered as an individual instance, to ten people coded as an average of 19 encounters. 50 eigenfaces were generated, and these were used to code learning images and a novel test set (image 20 for

each identity) which had not been used in constructing the averages. As in Study 1, a Mahalanobis distance metric was used in a nearest-neighbour match. This procedure was repeated 5 times, with level of familiarity rotated around identities. So, across the whole study, each identity was coded at each level of familiarity.

Results and Discussion

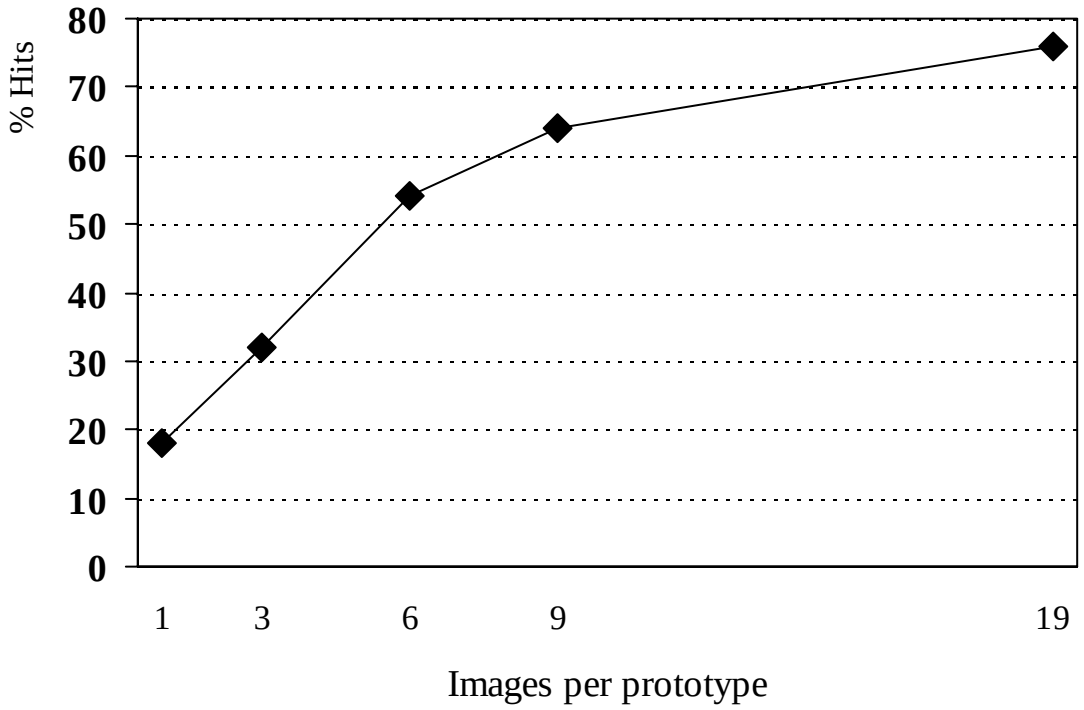

Figure 5 : Hit rate (% recognition) as a function of the number of images constituting each average representation.

Figure 5 shows the hit rate as a function of the number of images contributing to the average. There is clear improvement in hits as more images of each face are averaged together. Indeed there is a simple monotonic improvement in performance as the images contributing to each average increases. This seems to capture well the advantage for familiarity: the more encounters one has had with a person (coded here as the more images which contribute to the stored representation) the better is one's performance in recognizing a new image. This system is also beginning to perform at promising levels of performance for automatic recognition. Given the very variable input set, and unconstrained image capture conditions, a level of performance of 75% (for the 19-image averages) is encouraging. This performance is clearly not yet at a stage where it could be used practically for forensic identification or security purposes. However, the novel approach taken here, and the relatively poor performance of existing systems, suggests that this is an approach worth pursuing for applications-based as well as theoretical approaches to face recognition

Interim Summary

Taken together, these results seem to suggest that an abstractive system based on simple image averaging offers a useful way of thinking about face recognition. The very simplicity of the system is appealing: an arithmetic mean is perhaps the most obvious way to combine a set of examples, and it is therefore perhaps surprising that the system does so well. However, despite its simplicity, the averaging approach has a number of very appealing characteristics, which we will mention briefly here, before going on to examine human performance on these images.

One significant advantage of the averaging approach is that it appears to eliminate many of the surface characteristics of any specific image of a face. Each of the images in Figure 3a is the result of interactions between the face itself, the lighting conditions and the camera characteristics. Identifying the person therefore presents a very difficult problem, since a viewer who does not know the person does not know which visual properties are inherent to the person, which to the lighting and so forth. The image average, on the other hand, is not subject to this problem, since variations which are not characteristic of the person's identity are simply averaged-away. To see the most simple example of this, consider the fact that several of the images in Figure 3a are illuminated by noticeably directional lighting. When averaged together, this disappears, since it is not diagnostic of the person's identity. In fact, when coding knowledge of Tony Blair, one would almost certainly not want to incorporate lighting direction into one's visual representation of him, and the averaging process will automatically eliminate this. Note that this is an unintelligent strategy: there is no attempt to model the world of light, skin reflectance and camera properties. Some previous attempts to solve the automatic recognition problem have adopted this approach, but it is a very difficult problem to solve. The averaging process achieves the same aim, with a very simple technique.

In addition to the attractive nature of the representation itself, this technique seems to offer some promise in understanding the problem of face learning. Given the very marked differences between our ability with familiar and unfamiliar faces, this has been seen as a problem of processing shift. However, the averaging approach incorporates this shift naturally. Whether dealing with very familiar or less familiar faces, one is essentially matching incoming images to stored representations. What changes during the course of familiarization is that the stored representation becomes progressively refined. This refinement is not a progressive approximation to a particular likeness of a face, but a progressive *elimination* of all image properties which are not diagnostic of identity. In the early stages, when one is unfamiliar with a face, the viewer is forced into an image-matching strategy, because it is impossible to know which characteristics of a particular image are key to the identity of the person, and which are properties of the viewing and capture conditions. Indeed, subjects solving problems such as those in Figure 1, do seem to make simple image matches. In later stages, when robust averages have formed, the resulting representation captures information only relevant to identity, not to transient and superficial image properties.

We are arguing then, that the simple image average is a useful way to conceptualise stored representations of faces. The simulations above seem to suggest that the representation has some attractive properties, which we would want to incorporate into a model of human recognition. In the next section we will examine human perception of these images, and ask how they compare to perception of the constituent images which are used to build them.

Human face recognition

Our approach in this part of the paper is to investigate the human recognition of average faces. If these averages are, indeed, a good candidate for understanding our representations of familiar faces, then they should be well-recognised by observers. In order to test this, we use the same database of celebrities which was described in the sections above. For computer recognition, this was a convenient set, simply because it is possible to obtain many different images of famous people, taken across a large range of viewing conditions. In this part of the study, we exploit the fact that observers will know many of these people.

Figure 6: Some examples of averages, each formed from twenty shape-free instances. The top row shows shape-free images (used in simulations and in study 3. The bottom row shows these same images morphed to the average shape for that individual (used in studies 4 and 5). Names of people depicted are given in the appendix.

Figure 3b shows the effect of the shape-free manipulation on a particular famous face. It appears to us that some of these shape-free images preserve the person's identity rather well, and others less so. However, our informal observation is that the average of this person's shape-free images (e.g. Fig 3c) captures his identity well, consistent with Benson & Perrett's (1993) proposal. The top row of Figure 6 shows some further examples of average images. In these cases, averages were formed from all 20 images for each identity which were collected for Study 2.

At first, the fact that these images seem (to some extent) to preserve people's identity may seem surprising, because the shape to which they have been morphed is a simple face-shape template that retains none of the idiosyncratic characteristics of the originals. Later, we will consider the effects of putting shape back into the images (as in the bottom row of Figure 6), but in our first study of human perception, we will examine the simple shape-free averages. In particular, we ask whether the averaging technique does lead to more recognizable images, as more individual photos are used in their construction.

Study 3: Name verification to shape-free image averages

This study followed a name-verification procedure. Subjects were shown the name of a celebrity, followed by an image-average, which could be constructed from three, six or nine individual images. We measured their errors and reaction times to make this decision. The images were those of 24 of the celebrities used in Study 1. For that study, we constructed three 3-image averages, three 6-image averages, and a single 9-image average. These same images were used in the current experiment.

Method

Trials consisted of a celebrity's name presented at the center of the screen for 1500 msec, followed by a celebrity's face for 200 msec. Twelve volunteer subjects used speeded button-press responses (yes/no) to indicate whether or not the face matched the identity of the preceding name. On positive trials (50%), the face did match the name, and on negative trials (50%), a mismatch was presented..

Each subject carried out 3 blocks of 48 trials. In each block, they saw all 24 celebrities twice, once in a true trial and once in a false trial. The order of the entire sequence was independently randomized for each subject. Within a block, 8 faces comprised 3-image averages, 8 comprised 6-image averages, and the remaining 8 were 9-image averages. The n-averages were rotated around blocks, such that across the experiment, each celebrity was presented equally often as a 3-, 6- and 9-image average. The particular 3-image and 6-image averages used were held constant for each subject, but rotated about subjects, such that across the experiment each 3-image and 6-image average was used equally often.

Results and Discussion

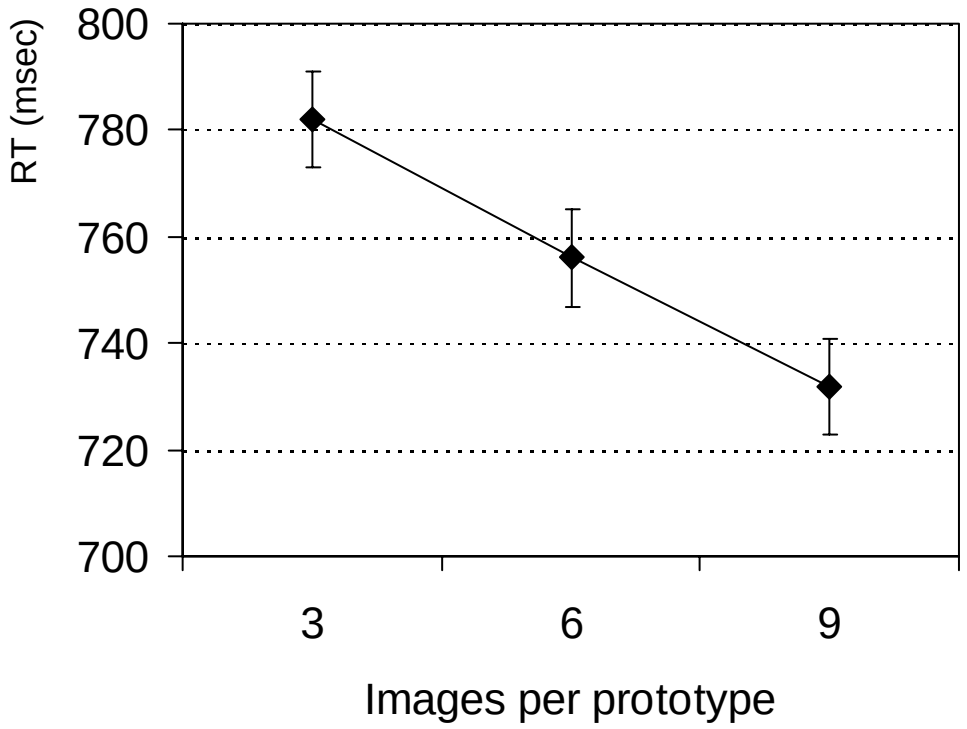

Figure 7: Mean correct verification RTs for averages constructed from 3, 6 or 9 contributing images (Study 3). Error bars are within subjects confidence intervals (Loftus & Masson, 1994).

Mean reaction time data for correct responses are shown in Figure 7. The pattern indicates that the number of images used to construct the average predicts recognition time. The 9-image averages are recognized fastest (732 msec), the 6-image averages (756 msec) more slowly, and the 3-image averages slowest (782 msec). (This pattern was confirmed by ANOVA, $F(2, 22) = 6.8$; $p < 0.01$). There was no systematic change in error rates across conditions (means: 18%, 22%, and 21% for the 3-, 6- and 9-image averages respectively, $F < 1$).

This pattern of results indicates that the averaging process provides a successively better image for recognition, as more and more individual photos are combined. Even though the shape of the images is held constant, and is not diagnostic of identity, the average (or template) of a particular person appears to improve with increasing sample size. This appears to provide a human experimental replication of studies 1 and 2, where a similar improvement was observed for artificial recognition. However, one difference is that we

have not presented a *single* shape-free image to subjects in the current experiment. This is partly because our observations of shape-free singleton faces, such as those in Figure 2b, do not seem to be highly recognizable. Note that the averages used in this experiment were recognised quite accurately, and so it is possible that eliminating shape-cues to identity has a particularly detrimental effect on individual instances of faces. Indeed, research using 3d models has suggested that recognition of identity relies on *both* texture and surface shape (O'Toole, Vetter & Blanz, 1999). In the next experiments we therefore consider the role of shape in human perception of averages.

Average shapes and average textures

While the pattern of data presented so far is encouraging for the averaging proposal, we are left with the problem that the image-averages we have used are devoid of informative (individuating) shape. Some previous studies of artificial face recognition have analysed shape and texture information independently (e.g, Calder et al, 2001; Hancock, et al, 1996). In these cases, using PCA, the texture information has been found to dominate recognition of identity. Furthermore, some studies of human face recognition have shown it to be highly tolerant of certain manipulations of the shape. For example, Hole, George, Eaves & Rasek (2002) demonstrated that familiar face recognition was completely unaffected by distorting the aspect ratio of photographs by up to 2:1, vertical to horizontal. On the other hand, research on the caricature effect suggests that manipulations of shape which emphasise idiosyncratic characteristics can improve identification (e.g., Rhodes, 1996). (Though note that these effects are most convincingly demonstrated when images are degraded or presented in a way which makes them difficult to recognise, such as using line drawings or brief presentations, e.g., Rhodes, Brennan & Carey, 1987; Lee, Byatt & Rhodes, 2000).

Despite situations in which shape appears not to dominate the recognition of identity, it is implausible that shape is simply ignored in human perception of faces. In the remaining part of this section on human recognition, we develop averages which incorporate both texture and shape information specific to each individual. We examine whether such averages can be recognized as well as individual images of the same person.

Figure 8 shows a diagrammatic representation of the shape-free procedure. In the first instance, a grid is dropped onto an image of a face. This is manipulated so that key points are identified in the image (corners of the mouth, of the eyes, and so forth). The image is then mapped onto a standard grid shape, using a morphing procedure. This delivers a shape-free face. However, the procedure also delivers the shape of the original face too, in the sense of identifying where the key points lay. In this way, Figure 8 demonstrates how the process of morphing to an average shape is a technique for separating two source of information in an image, the texture and the shape.

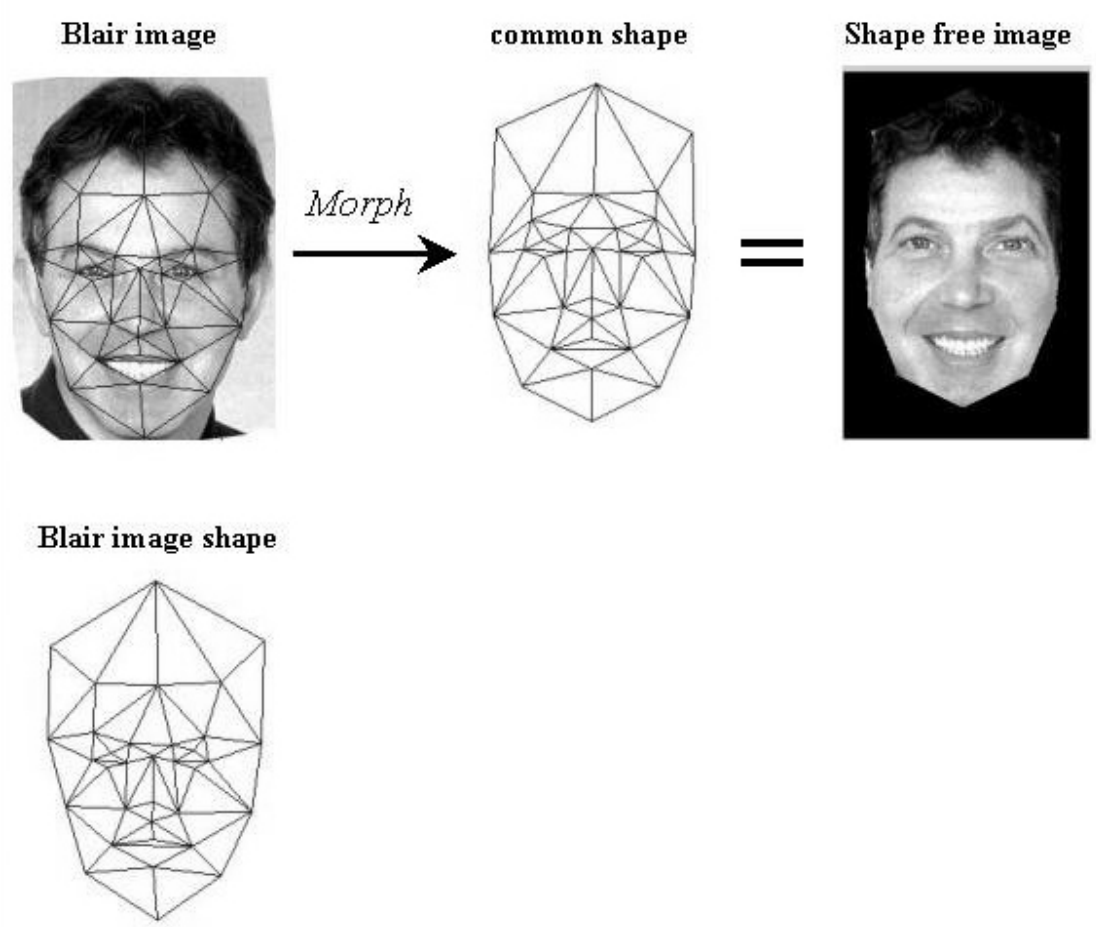

Figure 8: decomposition of a face image into shape and texture components

In studies 1 to 3, we have formed image averages by taking the mean of shape-free faces. However, it is also possible to derive the average shape of these images, simply by taking the mean xy positions for each grid point in the original image. It is therefore possible to derive an average shape for a set of images, as well as an average texture map. In the following experiments we use an average for each identity which is derived from both its average texture, and its average shape. This is computed by morphing the average texture for an individual, to that same person's average shape.

The bottom row of Figure 6 shows some example identity-averages constructed in this way. Some of the averages have been changed in quite a profound way (for example the first one) while others are changed less dramatically. All the faces in Figure 2 were constructed in this way, and our intuition is that ease of recognition is increased by comparison to shape-free images. To test this, we ran two further experiments, in which the identity-averages were compared directly to specific instances of faces.

Study 4: Name verification to identity-averages

In this experiment we use the same name-verification procedure as in Study 3. Following a name cue, a face appears, and this may be a specific image, or an identity average, as described above. Specific images were cropped to exclude background, and to give them an angular outer contour, as with the averages.

Method

16 volunteer subjects were recruited, all of whom reported normal, or corrected to normal vision. Each subject was presented with all fifty famous identities in each condition (as an average and as an instance). With fifty matched trials and fifty mismatched trials per condition, the experiment consisted of 200 trials in total, lasting approximately 15 minutes (including rest periods).

Trials consisted of a celebrity's name presented at the center of the screen for 1500 msec, followed by a celebrity's face for 200 msec. Subjects used speeded button-press responses (yes/no) to indicate whether or not the face matched the identity of the preceding name. On positive trials (50%), the face did match the name, and on negative trials (50%), a mismatch was presented. Order of stimulus presentation was randomized individually for each subject.

Results and Discussion

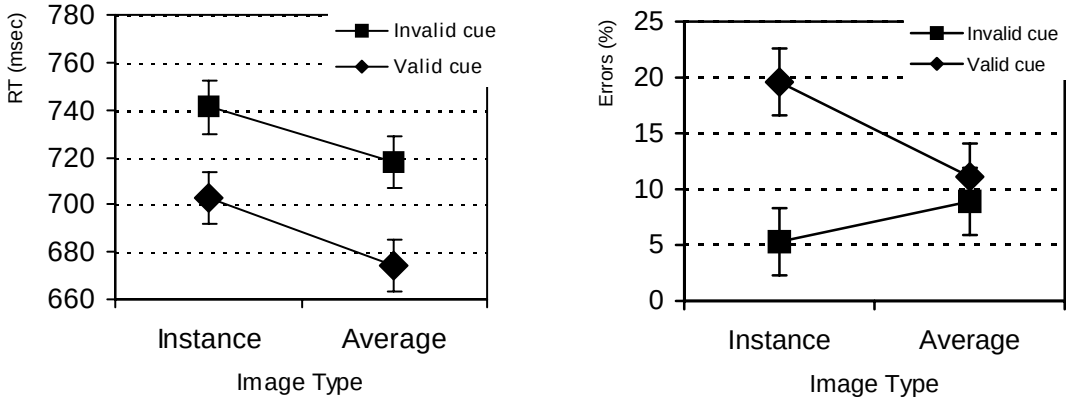

Figure 9: Mean RTs (Figure 9a) and Errors (Figure 9b) for study 4. Error bars are within subjects confidence intervals (Loftus & Masson, 1994), using pooled error variance (Loftus, 2004).

Figure 9 shows mean RT (9a) and error rates (9b) by condition. These data show that image averages are recognized faster than individual images. There is an effect of cue validity, as is usually the case in this procedure, but the important point to note is that both valid and invalid cues gave rise to the same advantage for averages over instances. (This pattern is confirmed by ANOVA showing reliable effects of instance/average, $F(1,15) = 6.7$, and valid/invalid, $F(1,15) = 24.4$, with no interaction, $F < 1$). Data for

errors is less clear, with the advantage for averages emerging only in the valid cue condition. (ANOVA confirms only a simple main effect of instance/average for valid trials, $F(1,15) = 20.3$, $p < 0.05$).

This seems to be quite compelling support for the notion that image averages are a good match to subjects' representations of familiar faces. Indeed, it is perhaps surprising that any representation can out-perform a specific instance of a face. However, one should note that subjects are viewing these images under unusual conditions. The name verification procedure used here employs a fast presentation rate, and subjects only see the images for 200ms. In the final experiment, we show the images to subjects under more normal viewing arrangements, and simply ask them to identify each person.

Study 5: Recognition of identity averages

Method

Previous testing revealed 10 identities from our set of 50 who were not well known to participants (INSERT FOOTNOTE 2). These 10 identities were removed from the set. For each of the remaining 40 faces, we used the same identity averages as used in Study 4 (derived from the average shape and average texture of 20 images of each individual). Two sets of instances (A and B) were chosen at random, such that each contained a single example of each individual. These sets of instances were compared (across subjects) in order to ensure that particular example photographs could not influence the overall results unduly (for example if one of the instances turns out to be a poor likeness of the person).

52 volunteer subjects were presented with 40 printed famous face images in a random order and were asked to identify each face by providing either the person's name (e.g. "Bill Clinton") or an individuating piece of semantic information (e.g. "the former president of the U.S.A."). Half of the faces were presented as identity instances and half as identity averages, so that each subject encountered each face in only one of these formats. Furthermore, half the subjects saw instances from set A, and half from set B. Presentation format and identity were counterbalanced across subjects so that over the course of the whole experiment, each face was presented as an average or an exemplar an equal number of times. Subjects were under no time constraints and were given as long as they wanted to complete the task.

Results and Discussion

Mean hit rates were 77% for instances and 81% for identity averages. These hit rates were, coincidentally, identical for instance set A and instance set B, and so no 2x2 analysis was necessary. A related means t test confirmed that identity averages were recognised reliably more often than instances ($t=3.57$; $p < 0.01$).

We have now demonstrated that the human perception of identity-averages is rather good. Using two different techniques, we have shown that identity averages can be preferred to individual images, even when these have been used in construction of the

average. We should, however, note that these results are *on average* results. That is to say, that identity-averages are on average better than individual images on average. It seems from looking at arrays of images of the same face (such as Figure 3a) that some individual images are simply better images of the person, in the sense of being more recognizable than others. The concept of a good or bad likeness in a photograph is commonplace in portraiture. It is possible, then, that some of the individual images we have used are better representations of the person than others. (We chose these at random from the 1000 images in our database, and not to be particularly good or bad likenesses). Note that the concept of a bad-likeness is only possible for an instance. The average, by comparison, can never be a bad likeness since it incorporates a large range of images. It is therefore perhaps unsurprising that, under these circumstances, a properly constructed average face can be a better representation than a randomly chosen instance. Of course, there may also be specific images which make particularly good recognition cues (for example well-known or iconic images of famous individuals), and so one cannot claim from this data that an average will always out-perform a specific instance. Nevertheless, we propose that this very simple representation, built using the simplest form of abstraction conceivable, seems to have the properties required of a robust representation for handling variability in input. In the final study, we return to further consideration of how this power may be exerting itself, using computer simulations to illustrate the issues.

Understanding the power of image averages

We have so far demonstrated that a very simple prototype system, based on image averaging, offers a promising representation for face recognition. A PCA-based system performs well with these averages, and appears to develop increasing expertise with increasing exposure (or number of prior encounters) with a face. Furthermore, there is some preliminary evidence that these image averages are perceived relatively accurately by human observers, particularly when shape is built-in to the averaging process. We now turn to an analysis of what might be underlying this effect, focusing particularly on the PCA system.

Study 6: Computational analysis of averaging

Why should storing an average face for each target perform better than several exemplars? Our explanation centers on the idea that averaging face images tends to remove artifacts due to difference in lighting, superficial image and camera characteristics, expression and small variations in pose, while consolidating information that is diagnostic of identity. (In what follows, we will group these non-identity variations under the general term ‘lighting’ for convenience.) To make this hypothesis explicit and to test whether it might account for the improvement seen, we now present a simulation based on an idealized “face space”. In this model, face images are held to exist in some multidimensional space. Some of the variations that we see are due to real differences between faces. Others are due to factors such as lighting and pose. The

problem that plagues recognition of faces is that the changes of the latter tend to outweigh the former, at least at the pixel level, so different faces seen under the same lighting can look more similar than the same face under different lighting. We model this with two sets of Gaussian random variables. The first set models the genuine differences between faces. The variance of these variables is relatively low but the average value for a given person will be non-zero, i.e. they will occupy a specific location in the "face space". The second set of variables models the variations due to artifacts such as lighting. These therefore have a relatively high variance, but a mean of zero, where zero means "average" lighting, whatever that might be.

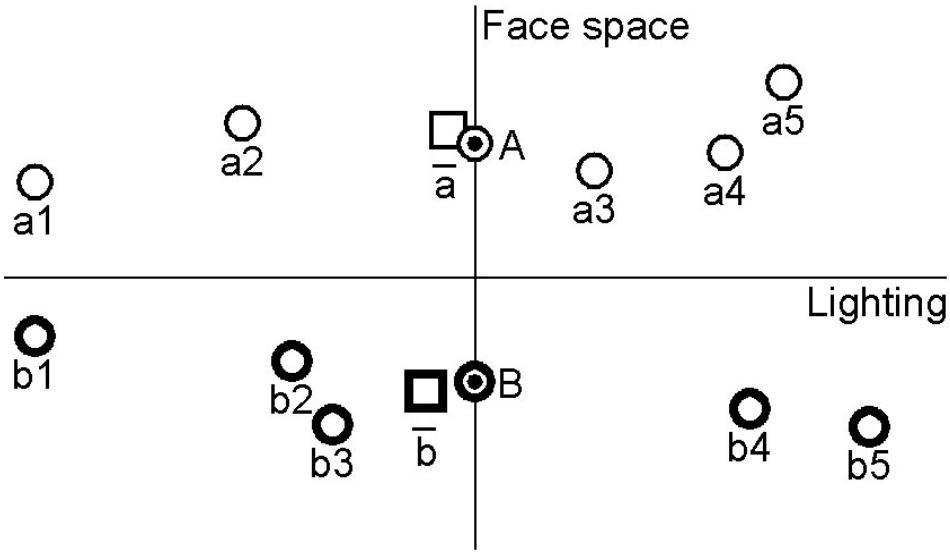

Figure 10. The space of face images is here represented by two dimensions, one capturing lighting and other extraneous changes, the other real differences in the appearance of faces. Two faces, A and B are shown, located on the origin of the lighting axis but separated in face space. Each face is represented by five face images, 1-5, which vary strongly on the lighting dimension and somewhat on the face space dimension (a face's natural variation over time). The averages of the exemplars are relatively near to the 'true' location of each face.

Figure 10 illustrates this idea in two dimensions, one for lighting and one for face space. Figure 10 represents a number of exemplar face images, which vary strongly on the lighting dimension, and their averages, which are closer to the origin. In the diagram, it can be seen that a simple exemplar model would fail for face B1, since it is nearer to A1 than any of the other B exemplars. However, it is nearer the B average than the A average. The diagram assumes Euclidean space: use of Mahalanobis distance, as in the work with real faces in the simulations above, will change the scales but not the interpretation.

The results of applying PCA to this model will depend strongly on whether averaging is carried out first. Without averaging, most of the variance lies along the lighting dimensions, so that is what PCA will pull out. The early components will code mainly lighting changes, irrelevant to identification. With averaging, much of the variance due

to lighting is eliminated and the PCA will be left with the face space dimensions, which are the ones that are interesting for recognition.

To assess how much effect this might have in practice, the following simulation was run. We assume 10 lighting and 10 face dimensions (entirely arbitrary figures that do not affect the form of the results). We simulate a situation in which there are fifty faces, each located at zero on the lighting dimensions and at a normally distributed random location in the face dimensions (standard deviation (sd) of the Gaussian distribution is 1). For each face, 10 simulated face images are generated, by adding a Gaussian random variable, sd=0.5, to each of the face dimensions and one of varying sd to each of the lighting dimensions. The lighting sd was varied between 0.3 and 1.5 to test the effect of this parameter. Within this model therefore, each canonical face is represented by a point at a random location in 10 dimensional face space and zero on the lighting axes, and individual face images are points somewhere in the 20 dimensional space of face and lighting components.

To test recognition, one example "face image" for each "face" was set aside to act as a probe. PCA was then run on the other 9 x 50 face images. We took the top 10 principal components as the basis set and transformed the 50 probe face images into this reduced space. (The effects of varying the number of components used will be discussed below). For each probe face the nearest neighbouring image, within the PC space, was identified, by Mahalanobis distance. This is the equivalent of the exemplar approach. To simulate the averaging approach, we first averaged the 9 exemplar face images for each face, to give 50 averages in the 20 dimensional model space. These averages were subjected to a new PCA. Each probe face image was then rendered into the new PC space and the nearest average face located, again by Mahalanobis distance. This whole process was repeated for each of the 10 different images for each face and then rerun a total of 10 times with new random face locations, to improve the regularity of the results.

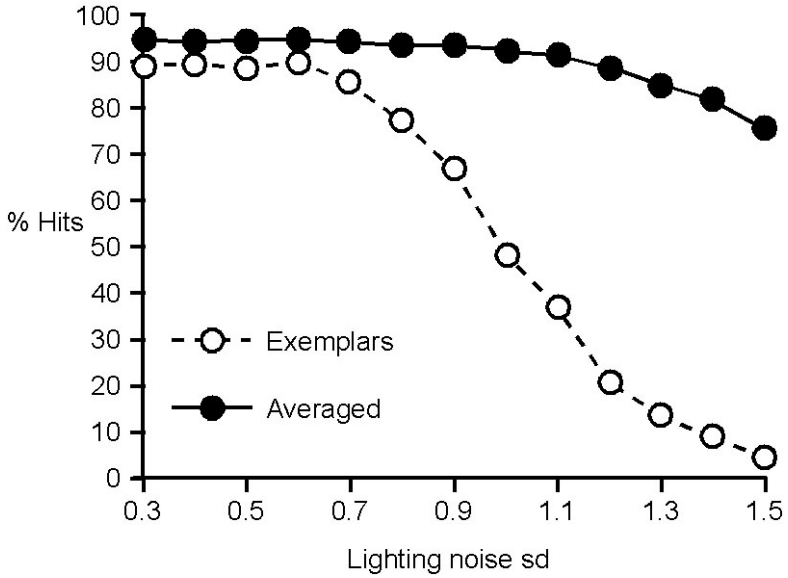

Figure 11: average hit rates for the average and exemplar methods, varying the amount of lighting noise.

Figure 11 shows the results of the simulation. As the sd of the lighting components increases, the average hit rate of the exemplar method rapidly falls. The increasing variability makes it more unlikely that the nearest other face image will be from the same face as the probe. Worse, the PCA will be affected by the variation in the lighting dimensions and code them preferentially, excluding the important face dimensions from the top 10 components. When the images are averaged before the PCA, the performance holds up much better. The amount of the improvement may initially seem surprising, since the probe images still carry all their lighting variation, and might therefore be expected to be a long way from the averages. However, this is where the gain of using PCA becomes apparent. Since the PCA was run on the averaged face images, the variance of the lighting dimensions was reduced. PCA therefore picks up the face dimensions and is simply not very sensitive to the variation in the lighting dimensions of a given exemplar used as a probe. Matching within the PCA space largely ignores the extraneous variations, leading to the big improvement in performance. The amount of improvement depends on the noise level, but it is evident that the model can account for the improvement seen with the real face images.

The improvement is also dependent on the number of components used. Here we used 10, which conveniently matches the number of face dimensions in the simulation. Use of fewer components reduces performance somewhat across the board, since useful information is thrown away. However, using too many has a strong effect on the averaged result, since these later components will code the lighting dimensions. Adding them in only increases the probability of false matches. We therefore anticipate that use of this simulation technique, linked to PCA on real averaged images, offers a potential source of information about the real dimensionality of face space. If future systems, based on realistically sized populations of face images, can achieve human levels of recognition accuracy, this will provide information about the number of dimensions needed to code faces within this simple linear scheme.

General Discussion

The work described here shows that a particular artificial face recognition system, based on PCA of image intensities, performs well with a very simple representation, derived from picture averages. The system is tested on a range of images, which are much more heterogeneous than normally used in reports of automatic recognition systems (Phillips et al, 2000; Zhao et al, 2003). Using this realistic range of superficial image characteristics, not normally noticeable until one sees them all together, as in Figure 3, there is immediate advantage for storing an average of learning images, over storing them all individually. Furthermore, better (more recognizable) averages are built from larger numbers of exposures. PCA has been proposed as a model of some aspects of human face processing (e.g., O'Toole et al, 1994; Burton et al, 1999) and the advantage it shows when using an abstract representation appears to offer potential for understanding human face recognition.

When one's task is to establish the identity of a face, superficial image cues such as contrast, illumination, lighting direction etc, must somehow be filtered out. However, an attempt to do this in a principled way, systematically accounting for each independently, and filtering accordingly, is a very difficult task. Instead, a simple image average automatically yields a face which is not subject to too great an influence from any of these factors, and we have simulated this above. Furthermore, as the average is taken from a larger and larger sample, the estimate of the "true mean" improves.

This simple proposal is attractive because it has potential to address a number of important issues in face recognition. First, it automatically provides an account of face learning. If face recognition can be understood as matching an image to a stored representation, then matching two images of an unfamiliar face will essentially be an image-matching (rather than a face-matching) task. Bruce et al, (1999, 2001) and Hancock et al (2000) suggest that this is exactly the strategy used in unfamiliar face matching, in contrast to a more abstractive approach in which some canonical knowledge of face variation is recruited. In order to become an expert with a familiar face, and be able to recognise it over an increasing range of visual conditions, one simply needs to improve one's representation. Under this scheme, matching familiar faces is the same process as matching unfamiliar faces. The huge difference in one's facility to do this arises simply because of a much better target against which to match familiar faces. Notice that this proposal is not inconsistent with previous work suggesting that internal features become more important for recognizing a face as it becomes more familiar (Ellis et al, 1979; Young et al, 1985). If, over a range of exemplars, it is the internal features which remain constant, while external features such as hairstyle change, then it is the internal features which will be preserved automatically by the averaging process.

This proposal also allows a way of capturing the difficulty of recognizing people over a very large range of circumstances, including changes through life. Viewers of a certain age, tested in psychology laboratories, have no difficulty recognizing a photograph of Paul McCartney taken any time between 1960 and 2005. Photographs of McCartney at 20 look very different from photographs at 60 years old, and yet people recognise each with ease. One might anticipate that this could only be achieved by storing separate representations of McCartney, one for each age. However, this simple image averaging technique preserves precisely those aspects of a face's identity which remain constant over the range of images which constitute it. The technique eliminates not only superficial properties due to light and photographic equipment, but also properties due to age, mood, health, and so forth. Notice that the resulting average of McCartney need not necessarily look like a photograph of McCartney. To be sure, it will be free of wrinkles, detailed complexion, and many of the other properties of images (see Figure 2). However, it need only act as a McCartney filter to be effective. So, it need only be more like any input image of McCartney, than is any of the other stored averages, in order to work as an effective and efficient representation.

This proposal appears to have a number of characteristics which render it a potentially interesting form of representation for understanding face recognition. Building on a line of theorizing using putative Face Recognition Units (Bruce & Young, 1986), we have

offered one way of implementing such units. Figure 2 provides a way to visualise this proposal for a set of known faces. In offering it, we certainly do not wish to preclude future instance-based systems. We have not, of course, tested the whole range of possible exemplar-based formulations, and it is possible that a future system will hold just as much promise as the one presented in this paper. However, we hope that we have demonstrated that this particular prototype-based formulation, derived from a very simple averaging technique, offers considerable potential for future research addressing a wide range of face recognition problems.

Concluding remarks

Although we have offered an outline solution to the problem of face representation, there are clearly very many issues outstanding. In this final section we discuss some of these in the hope that we can be as clear as possible in articulating what is and is not claimed for this proposal.

1. Who decides which faces are averaged together?

The proposal we offer here relies on a supervised learning technique. When a new image of Tony Blair is perceived, it is necessary to know who it is, in order to incorporate it into the average of the correct person. It is clear that there are very many occasions in which person recognition is not only based on identification of a face, and one has support from many other sources in order to make the identification (for example, voice, clothing and social context). Furthermore, in social interaction (or simple observation), one is given very many examples of how that person's face may appear, moment by moment, across changes due to head position, expression and speech. Under all these circumstances there is very strong top-down support for deciding which representation to update with a new instance. Of course, there will be occasions on which these supporting structures are absent, for example if one were to see a familiar person in an unexpected place. Under these circumstances, we would expect the system sometimes to make a mistake. However, note that this is exactly the situation in which human perceivers make mistakes. The clear prediction from our proposal is that these mistakes would be more common for less-well known individuals than for people who are very well known to the system, and who have the more robust representations coded. This seems to be a rather uncontentious prediction to make.

2. Is a single full-face template all that is needed for familiar face recognition?

Although we have only provided studies of full-face images, there is good evidence that this is insufficient for a robust representation of familiar people. For example, it has been known for many years that certain views (and particularly 3/4 views) seem to be particularly well recognised, by comparison to full face or profile views, though this effect is moderated by familiarity and learning conditions (e.g., Bruce, Valentine & Baddeley 1987; Logie, Baddeley & Woodhead, 1987; Liu & Chaudhuri, 2002). However, evidence from a range of sources suggests that effects of view dependence need not necessarily arise from a generalisable ("rotatable") representation. Instead, a small number of canonical views (for example full face, three quarter and profile) can be used to generalize to other intermediate views without significant decrement in recognition

performance (see e.g., Hill, Schyns & Akamatsu, 1997; Perrett et al, 1985, 1998 ; Logothetis, Pauls & Poggio, 1995). This position is consistent with the averaging proposal outlined here. In order to store a fully robust representation of a known face, it seems likely that one will have to store, separately, averages of that person's full face, 3/4 view, and profile. However, the means by which this is achieved could be the same as those outlined above, without loss of generalization. We are therefore not proposing that full face is a sufficient representation for face recognition, but that a small number of discrete viewpoints will be necessary to generalize the proposal.

3. The grid shape is manipulated by hand, should this be automatic?
In the studies described here, all key points for grid placement were found by hand. In one sense this detracts from the claim that this proposal may be useful in automatic face recognition. However, our averaging proposal relies on face standardization, and the mechanism for achieving a standard shape must work reliably in order for averages to be useful. The proposal therefore requires that there is some analogue of standardization in human face perception. This seems to be a reasonable notion, though we have not offered a mechanism for achieving it. In fact, it seems likely that any standardization mechanism acts independently of identification processes. Certainly, subjects who fail in tasks such as shown in Figure 1, have no difficulty in locating the key features of the face, as required for grid placement. While it would be possible to automate the standardization process using any one of a number of computer-vision techniques (e.g. techniques related to those proposed by Blanz & Vetter, 1999), such a mechanism would be independent of the current specific proposal. Furthermore, any such system would inevitably introduce further errors. We have therefore chosen to present a system uncontaminated by such errors, while acknowledging that this extra component would be needed for any future practical deployment of the scheme.

4. What are the relative contributions of shape and texture to identification?
The results of our artificial face recognition studies, as well as previous reports in the literature (e.g., Burton et al, 2001, Calder et al, 2001), show that PCA performs quite well with shape-free images. However, this does not mean that shape is unimportant in recognising faces. Note that in the shape-free versions of faces, information about shape is nonetheless present. So, for example, the pattern of pixel intensities for a shape-free chin, will be different depending on whether the original was a big or a small chin. Since the PCA is not tuned to any particular face-shape (i.e. the shape we choose is essentially arbitrary for the computer analysis) this extra information is available for use in the performance. However, as we have shown in the studies above, human recognition of averages is rather good with shape included. It would therefore be worthwhile topic of future study to ask how these two sources of variation combine. O'Toole et al (1999) have studied combination of 3d shape and texture information, finding both to be important for identification, however comparable studies do not yet exist for 2d stimuli. Our initial observations are that shape provides good support for face recognition in this situation, but is not a dominant cue. Raw grid information, such as shown in figure 8, is never recognizable. We have also tried to morph average face textures to an individual's shape, and again this never results in recognizable faces.

However, we should note that we have chosen to use a very simple grid. It is possible that future research, using a grid with more fine-scale resolution, would pick up independent effects of shape on face recognition. This will be a topic for future research.

References

Bachmann, T. (1991). Identification of spatially quantised tachistoscopic images of faces: How many pixels does it take to carry identity? European Journal of Cognitive Psychology, 3,b87-103.

Benson, P.J., & Perrett, D.I. (1993). Extracting prototypical facial images from exemplars. Perception, 22, 257-262.

Beymer, D. (1995). Vectorizing Face Images by Interleaving Shape and Texture Computations. AI Lab, Memo 1537, MIT, Sept. 1995.

Blanz, V. & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. SIGGRAPH 1999, 187-194. New York: ACM Computer Society Press.

Bonner, L., Burton, A.M. & Bruce, V. (2003). Getting to know you: How we learn new faces. Visual Cognition, 13, 527-536.

Brédart, S., Valentine, T, Calder A.J., & Gassi L. (1995). An interactive activation model of face naming. Quarterly Journal Of Experimental Psychology, 48A, 466-486.

Bruce, V. (1994). Stability from variation: The case of face recognition. The M.D. Vernon Memorial Lecture. Quarterly Journal of Experimental Psychology, 47, 5-28.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P., Burton, A.M. & Miller, P. (1999). Verification of face identities from images captured on video. Journal of Experimental Psychology: Applied, 5, 339-360.

Bruce, V., Henderson, Z., Newman, C. & Burton, A.M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. Journal of Experimental Psychology: Applied, 7, 207-218.

Bruce V., Valentine, T. & Baddeley A. (1987). The basis of the 3/4 view advantage in face recognition. Applied Cognitive Psychology, 1, 109-120.

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

Burton, A.M., Bruce, V. & Hancock, P.J.B. (1999). From pixels to people: a model of familiar face recognition. Cognitive Science, 23, 1-31.

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

Burton, A.M., Miller, P., Bruce, V., Hancock, P.J.B. & Henderson, Z. (2001). Human and automatic face recognition: a comparison across image formats. Vision Research, 41, 3185-3195.

Burton, A.M, Wilson, S., Cowan, M & Bruce, V. (1999). Face recognition in poor quality video: evidence from security surveillance. Psychological Science, 10, 243-248.

Burton, A.M., Young, A.W., Bruce, V., Johnston, R.A. & Ellis, A.W. (1991). Understanding covert recognition. Cognition, 39, 129-166.

Calder, A.J., Burton, A.M., Miller, P., Young, A.W. & Akamatsu, S. (2001). A principal component analysis of facial expressions. Vision Research, 41, 1179-1208.

Clutterbuck. R. & Johnston R.A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. Perception, 31, 985-994.

Cottrell, G.W., Branson, K. & Calder, A.J. (2002) Do expression and identity need separate representations? In Proceedings of the 24th Annual Cognitive Science Conference, Fairfax, Virginia. Mahwah: Lawrence Erlbaum

Craw, I. (1995). A manifold model of face and object recognition. In T. Valentine (Ed.) Cognitive and computational aspects of face recognition. London: Routledge.

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.

Dailey, M.N., Cottrell, G.W., Padgett, C. & Adolphs, R. (2002). EMPATH: A neural network that perceives and categorizes facial expressions. Journal of Cognitive Neuroscience, 14, 1158-1173.

De Haan, E.H.F., Young, A.W., & Newcombe, F. (1987). Face recognition without awareness. Cognitive Neuropsychology, 4, 385-415.

Ellis, A.W., Flude, B.M., Young, A.W. & Burton, A.M. (1996). Two loci of repetition priming in the recognition of familiar faces. Journal of Experimental Psychology: Learning Memory and Cognition, 22, 295-208.

Ellis, A.W. Young, A.W., & Flude, B. (1990). Repetition priming and face processing: Priming occurs within the system that responds to the identity of a face, Quarterly Journal of Experimental Psychology, 42A, 495-512.

Ellis, H.D. (1986) Processes underlying face recognition. In R. Bruyer (Ed.), The neuropsychology of face perception and facial expression. Hillsdale, NJ: Erlbaum.

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

Furl, N., Phillips, P.J., & O'Toole, A.J. (2002). Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis Cognitive Science, 26, 797-815.

Gonzalez, R.C. & Woods, R.E. (2002). Digital Image Processing, 2nd Edition. Prentice Hall.

Hancock, P.J.B., Burton, A.M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. Memory & Cognition, 24, 26-40.

Hancock, P.J.B., Bruce, V. & Burton, A.M. (2000). Recognition of unfamiliar faces Trends in Cognitive Science, 4, 330-337.

Hanley, J.R. (1995). Are names difficult to recall because they are unique - a case-study of a patient with anomia. Quarterly Journal Of Experimental Psychology, 48A, 487-506.

Hanley, J.R. & Turner, J.M. (2000). Why are familiar-only experiences more frequent for voices than for faces? Quarterly Journal of Experimental Psychology, 53A, 1105-1116.

Harmon, L.D. (1973). The recognition of faces. Scientific American, 227, 71-82.

Harmon, L.D. & Julesz, B. (1973). Masking in visual recognition: Effects of two-dimensional filtered noise. Science, 180, 1194-1197.

Hay. D.C. & Young, A.W. (1982). The human face. In A.W. Ellis (Ed.) Normality and pathology in cognitive functions. 173-202. London: Academic Press.

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

Hole, G.J., George, P.A., Eaves K., & Rasek, A. (2002). Effects of geometric distortions on face recognition performance. Perception, 31, 1221 – 1240.

Kemp, R., Towell. N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. Applied Cognitive Psychology, 11, 211-222.

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

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

Lee, K., Byatt, G. & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: testing the face-space framework. Psychological Science, 11, 379-85.

Loftus, G.R. (2004). Analysis, Interpretation, and Visual Presentation of Experimental Data. In H. Pashler & J. Wixted (Eds.) Stevens' Handbook of Experimental Psychology, Volume 4, Methodology in Experimental Psychology. London: Wiley.

Loftus, G.R. & Masson, M.E.J. (1994). Using confidence intervals in within-subjects designs. Psychonomic Bulletin and Review, 1, 476-490.

Logie, R.H., Baddeley, A.D., Woodhead, M.M. (1987). Face recognition, pose and ecological validity. Applied Cognitive Psychology, 1, 53-69.

Logothetis, N.K., Pauls, J. & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Current Biology, 5, 552-563.

Moghaddam, B., Nastar, C., & Pentland, A. (1996). Bayesian face recognition using deformable intensity surfaces. Proceedings of Computer Vision and Pattern Recognition '96, 638-645.

Morton, J. (1969), Interaction of information in word recognition, Psychological Review, 76, 165-178.

Nosofsky, R.M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning Memory and Cognition. 14, 700-708.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception and Performance, 17, 3-27

O'Donnell, C., & Bruce.V. (2001). Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. Perception, 30, 755-764.

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.

O'Toole, A.J., Vetter, T., & Blanz, V. (1999). Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing. Vision Research, 39, 3145-3155.

Perrett. D.I., Oram, M.W. & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. Cognition, 67, 111-145.

Perrett. D.I., Smith, P.A.J.,, Potter, D.d., Mistlin, A.J., Head, A.S., Milner, A.D. & Jeeves, M.A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. Proceedings Of The Royal Society Of London Series B-Biological Sciences, 223, 293-317.

Philips, P.J., Grother, P., Michaelis, R.J., Blackburn, D.M., Tabassi, E., & Bone, J.M. (2003). Face recognition vendor test 2002: Evaluation report NISTIR 6965. Available online at http://www.frvt.org

Phillips, P.J., Moon, H., Rizvi, S.A. & Rauss, P. (2000). The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Patern Analysis and Machine Intelligence, 22, 1090-1104.

Rhodes, G. (1996). Superportraits: Caricatures and Recognition. Hove, UK: Psychology Press.

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

Schweinberger, S.R. (1996). How Gorbachev primed Yeltsin: Analyses of associative priming in person recognition by means of reaction times and event-related brain potentials. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 1383-1407.

Schweinberger, S.R., Herholz. A., & Stief, V. (1997). Auditory long-term memory: Repetition priming of voice recognition. Quarterly Journal of Experimental Psychology, 50A, 498-517

Schyns, P.G., Oliva, A. (1997). Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition Perception, 26, 1027-1038.

Sergent, L. (1986). Microgenesis of face perception. In H.D. Ellis. M.A. Jeeves, F. Newcombe, & A.W. Young (Eds.), Aspects of face processing. Dordecht: Martinus Nijhoff.

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

Valentin, D., Abdi, H. & O'Toole, A. (1994). Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches. Journal of Biological Systems, 2, 413-423.

Vetter, T. and Troje, N. (1995) Separation of texture and two-dimensional shape in images of human faces. In: Sagerer, G., Posch, S. and Kummert F., Mustererkennung 1995, Reihe Informatik aktuell, pp. 118-125, Springer Verlag.

Wiskott, L., Fellous, J-M., Kruger, N. & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis & Machine Intelligence, 17, 775-779.

Yambor, B. Draper and R. Beveridge (2002). Analyzing PCA-based Face Recognition Algorithms: Eigenvector Selection and Distance Measures in , H. Christensen & J. Phillips (Eds.), Empirical Evaluation Methods in Computer Vision. World Scientific Press, Singapore.

Young, A.W. & Bruce, V. (1991). Perceptual categories and the computation of "Grandmother". (1991). European Journal of Cognitive Psychology, 3, 5-49.

Young, A.W. & Burton, A.M. (1999). Simulating face recognition: implications for modelling cognition. Cognitive Neuropsychology, 16, 1-48.

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

Young, A.W., Hellawell, D., & DeHaan, E.H.F. (1988). Cross-domain semantic priming in normal subjects and a prosopagnosic patient. Quarterly Journal of Experimental Psychology, 40A, 561-580.

Zhao, W., Chellappa, R., Phillips, P.J. & Rosenfield, A. (2003). Face recognition: A literature survey. ACM Computing Surveys, 35, 399-458.

Footnotes

Footnote 1: We have repeated this study using a Euclidean metric. Consistent with the literature, we found that Mahalanobis distance matches were better than Euclidean matches in every version of the system. Since a system based on Euclidean matching is not a serious candidate for this type of recognition problem, we have therefore not presented data on this manipulation, though it is available from the authors on request.

Footnote 2: The appendix lists all 50 celebrities in our database, and averages of these people are shown in Figure 2. Although all are famous, some are better known by certain age groups than others. The student sample for study 6 was not familiar with some of the older celebrities.

Appendix: People depicted in figures

Figure 1: in line-up 1a, the target is number 3; in line-up 1b the target is not present.

Figure 2: From top left in rows: Al Pacino, Bill Clinton, Brad Pitt, Cameron Diaz, Catherine Z Jones, Cher, Cherie Blair, Clint Eastwood, David Beckham, David Bowie, Elvis Presley, Ewan McGregor, George Bush, Geri Halliwell, Gwyneth Paltrow, Harrison Ford, Jack Nicholson, Jennifer Anniston, Jennifer Lopez, John Travolta, Julia Roberts, Keanu Reaves, Kevin Spacey, Kylie Minogue, Leo di Caprio, Liz Hurley, Madonna, Margaret Thatcher, Marilyn Monroe, Meg Ryan, Mel Gibson, Michael J Fox, Michelle Pfeiffer, Natalie Portman, Nicholas Cage, Paul McCartney, Princess Diana, Russell Crowe, Sarah J Parker, Sarah M Gellar, Sean Connery, Sharon Stone, Susan Sarandon, Sylvester Stallone, Tom Cruise, Tony Blair, Uma Thurman, Victoria Beckham, Vinnie Jones, Winona Ryder. (Note that celebrities were chosen as being famous to a British audience.)

Figure 6: From left to right: John Travolta, Susan Sarandon, Sylvester Stallone, Leo di Caprio