100% Accuracy in Automatic Face Recognition

R. Jenkins* and A. M. Burton

National security and crime prevention often depend on our ability to establish the identities of individuals and check that they are whom they claim to be. This proof of identity is frequently achieved by comparing the individual's appearance to a photo-identification document such as a passport. Although there are now a number of automatic face-recognition devices available, none can cope with the kind of image variability encountered in the real world (*1*). Even in relatively constrained settings performance is far from perfect, and current accuracy levels would translate to thousands of errors in any large-scale system (e.g., transport security). The only system that can reliably cope with real-world image variability is a human observer who is familiar with the faces concerned (*2*). We have recently proposed that human familiarity with a particular face can be modeled by a process of image averaging (*3*), whereby different photos of that face are merged to form a single image (*4*). Here, we show that image averaging can also greatly improve performance of an automatic face-recognition system.

FaceVACS (Cognitec Systems GmbH, Dresden, Germany) (*5*) is an industry standard face-recognition system that has been widely adopted [e.g., SmartGate (Australian Customs Service) at Sydney Airport]. For these studies, we used an online implementation of FaceVACS at the genealogy Web site MyHeritage (My Heritage Limited, Tel Aviv, Israel) (*6*). We had no control over the algorithm used by this system or the database of known faces. The database contained 31,077 photographs of famous faces, comprising an average of nine different photos for each of 3628 celebrities. These photographs were collected from diverse sources and were taken over several decades with various cameras. They are thus highly variable in their quality and cover a wide range of lighting conditions, facial expressions, poses, and age. Users of the Web site upload their own face images, and the system returns the closest matching photograph from its database.

We fed this system photographs from our own famous face database (*5*) in order to assess its accuracy on images of real-world variability. When the identity of the returned photograph matched that of the uploaded image, we recorded a hit. Otherwise, we recorded a miss. Our probe database consisted of 500 images, comprising 20 different photographs for each of 25 male celebrities who were also in the online database. Forty-one of the probe images were identical to images in the online database, and these were excluded from the analysis. The overall hit rate for the remaining 459 different images was 54%. The hit rate for individual faces varied according to the number of images of that face that were in the online database. Performance ranged from 16% correct when seven images were stored to 89% correct when 28 were stored. We next sought to establish whether image averaging could improve overall performance. For each test identity, we created a new image by averaging together the 20 images of that person in our probe database (Fig. 1). Note that the online database and the matching algorithm remained the same; the only change from the first study was that we merged the probe images to create an average image for each face (fig. S1). Surprisingly, this simple process raised the hit rate from 54% to 100%. This is unprecedented for such varied images.

It is possible that the averages were well recognized simply because they incorporated some recognizable photos. To rule out this possibility, we constructed a new set of averages using only those photographs that were unrecognized in the first study. That is, we fixed baseline performance at 0%, so that any improvement above 0% could be attributed solely to the averaging process. Applying image averaging to these missed items raised the hit rate from 0% to 80%.

Our findings show that the simple process of image averaging can dramatically boost automatic face recognition. We demonstrated this improvement with a commercially available algorithm and an online face database over which we had no control. We suggest that image averaging enhances performance by stabilizing the face image. With standard photographs, the match tends to be dominated by aspects of the image that are not diagnostic of identity (e.g., lighting and pose). Averaging together multiple photographs of the same person dilutes these transients while preserving aspects of the image that are consistent across photos. The resulting images capture the visual essence of an individual's face and elevate machine performance to the standard of familiar face recognition in humans. It would be technically straightforward to incorporate an average image into identification documents. Doing so would greatly reduce the incidence of face-recognition errors and raise the prospect of a viable automatic face-recognition infrastructure.

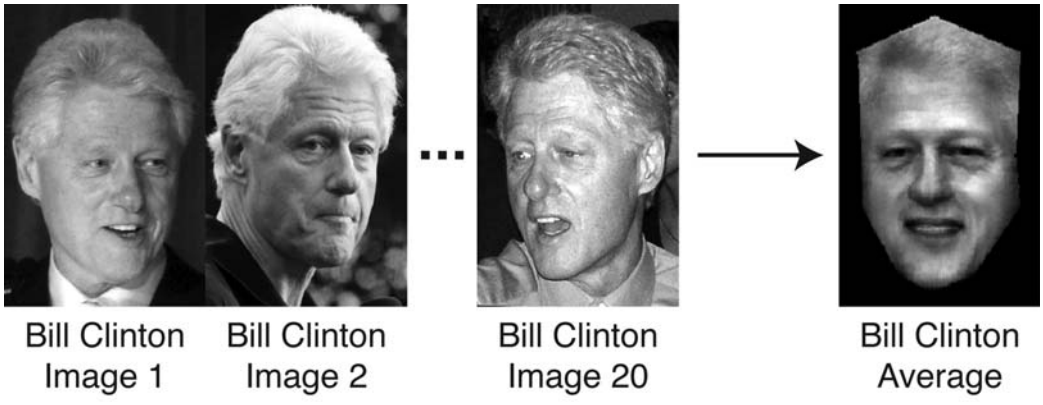


Fig. 1. Example photographs of Bill Clinton and their average (**right**). [Image 1, photo by Marc Nozell (www.flickr.com/photos/marcn/534512066); image 2, photo by Roger Goun (www.flickr.com/photos/sskennel/829574139); image 20, photo by Nelson Pavlosky (www.flickr.com/photos/skyfaller/26752190). All photos were used under a Creative Commons license.] Different pictures of a single face can vary enormously, making automatic recognition difficult. Averaging together multiple photos of the same face stabilizes the image, improving performance dramatically.

References and Notes

1. W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfield, *Assoc. Comput. Mach. Comput. Surv.* **35**, 399 (2003).
2. P. J. B. Hancock, V. Bruce, A. M. Burton, *Trends Cogn. Sci.* **4**, 330 (2000).
3. A. M. Burton, R. Jenkins, P. J. B. Hancock, D. White, *Cogn. Psychol.* **51**, 256 (2005).
4. F. J. Galton, *Nature* **18**, 97 (1878).
5. Materials and methods are available on *Science* Online.
6. www.myheritage.com

7. We thank G. Cohen at MyHeritage for information about their database. This research was funded by the Economic and Social Research Council, UK.

Supporting Online Material

www.sciencemag.org/cgi/content/full/319/5862/435/DC1
Materials and Methods
Fig. S1
References and Notes

23 August 2007; accepted 15 November 2007
10.1126/science.1149656

Department of Psychology, University of Glasgow, Glasgow G12 8QQ, UK.

*To whom correspondence should be addressed. E-mail: rob@psy.gla.ac.uk

Supporting Online Material for

100% accuracy in automatic face recognition

Rob Jenkins and A. Mike Burton

Materials and methods

FaceVACS. Technical details and current applications of the *FaceVACS* system are available from the developer's website *(1)*. Test data and performance comparisons with other automatic face recognition systems are published regularly *(2)*.

Probe images. Our probe database consisted of 500 photographs of famous faces collected from the internet. Each photo showed the full face without occlusion in roughly frontal aspect (tolerance of approximately 30˚). The images varied widely in terms of lighting conditions, overall size, focus, and general picture quality, and covered a natural range of variation in facial expression, pose, age, and hairstyle. Each image was converted to grayscale, and rotated in the picture plane to bring the eyes to within 5˚ of horizontal. The face region was then resized and cropped to 190 x 285 pixels, preserving the original aspect ratio. Image contrast was enhanced using the Auto Contrast function of Adobe PhotoShop with default settings.

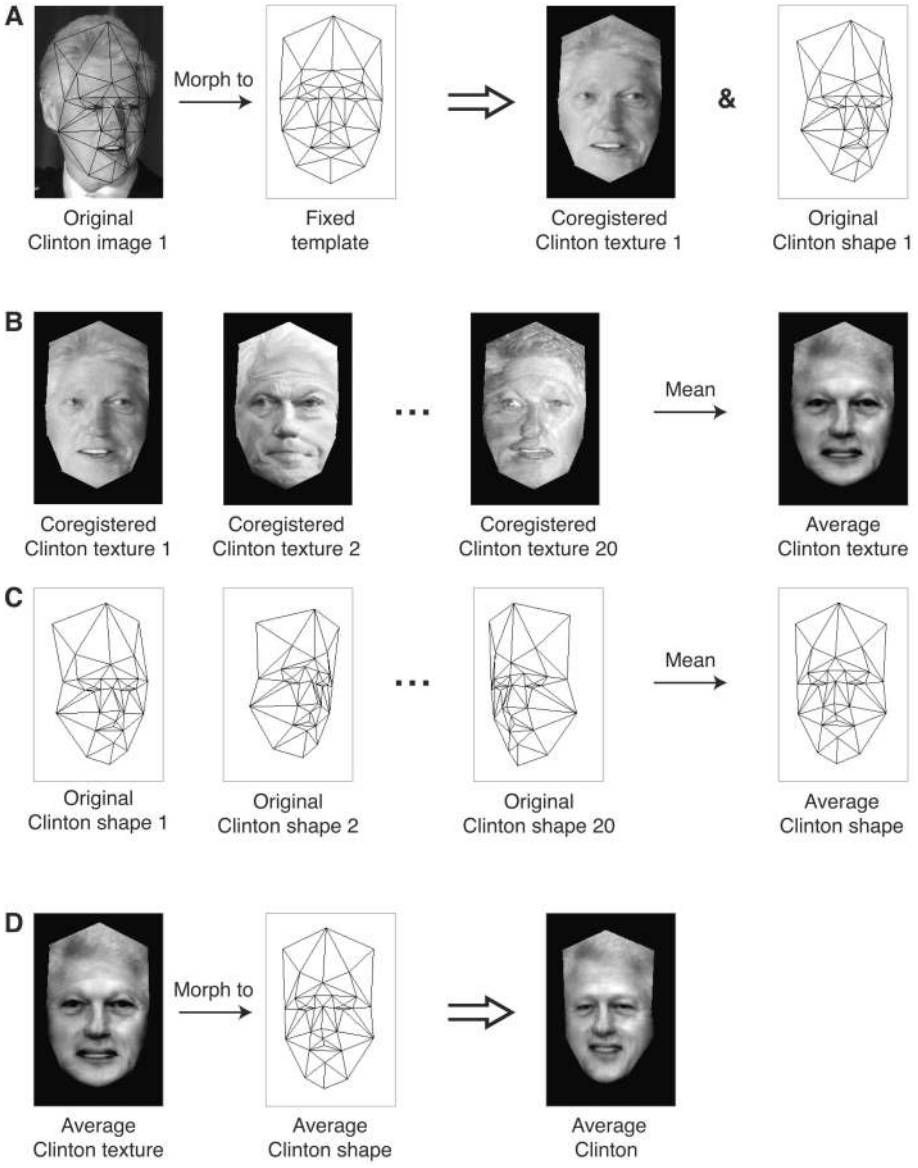


Fig. S1. Image averaging procedure. (**A**) Facial shape in each image was captured by recording the xy-coordinates of 34 facial landmarks (e.g. corners of the eyes, tip of the nose). The images were then co-registered by morphing them to a fixed 34-point template using bi-cubic interpolation. For each face we derived (**B**) the average texture from 20 co-registered images, by calculating the mean grayscale value at each pixel, and (**C**) the average shape of the corresponding unregistered images, by calculating the mean xy-coordinates of each facial landmark. (**D**) We then morphed each person's average texture to their average shape to produce the average image of their face. Image contrast was enhanced as for the standard photographs.

Comment on "100% Accuracy in Automatic Face Recognition"

Weihong Deng,* Jun Guo, Jiani Hu, Honggang Zhang

Jenkins and Burton (Brevia, 25 January 2008, p. 435) reported that image averaging increased the accuracy of the automatic face recognition to 100% and thus could be applied to photo-identification documents. We argue that the feasibility of image averaging on identification documents is not fully supported by the presented evidence.

In automatic face recognition, a gallery of facial images is first enrolled and coded for subsequent searching. A probe image is then obtained and compared with each encoded face in the gallery, and a recognition is noted when a suitable match occurs. In a recent study, Jenkins and Burton (*1*) used the photographs of celebrities as probe images to measure the hit rate of the FaceVACS (Cognitec Systems GmbH, Dresden, Germany) face-recognition system used by the genealogy Web site MyHeritage (*2*). Merging the probe images to create an average image for each celebrity raised the overall hit rate for the probe database from 54 to 100%. The authors therefore concluded that the process of image averaging could dramatically boost automatic face recognition and inferred that incorporating average face images into identification documents would greatly reduce the incidence of face-recognition errors.

As Jenkins and Burton suggest in (*1*), it is possible that 100% accuracy was achieved simply because the image averages incorporated some recognizable photos. To allay that concern, the authors reported that a new set of averages using only those photographs that were unrecognized in their first study raised the hit rate from 0 to 80%. They thus reasoned that the improved accuracy could solely be attributed to the averaging process. However, the improvement on the hit rate could be partly attributed to the manual facial registration [see supporting online material for (*1*)] before averaging, which accurately rectified the facial appearance so that all the probe faces were aligned in a standard frontal and upright posture and enclosed by a uniform background. By largely reducing the image variability, the image registration procedure might transform some unrecognized photos into recognizable faces (*3*). Moreover, the standard registered faces might facilitate the automatic face finding and normalization process of the tested algorithm, which may have also boosted the hit rate (*4*). It is thus possible that the registration technique assisted the image averaging to boost the hit rate to a higher level.

Jenkins and Burton correctly suggested that image averaging enhanced the performance by stabilizing the face image. However, the interpretation of this fact was overextended. The conclusion that including average images on identification documents would reduce recognition errors lacks sufficient evidence, especially because it is not an equivalent task to the experiments that were carried out. Specifically, the experiments in (*1*) used the online database as the gallery and the average images as the probe, and the online recognition system only returned the closest matching photo from its database. If the identity of the returned photo matched that of the average image, it was recorded as a hit. Using this methodology, even the 100% hit rate could only ensure that, for each test identity, the system successfully matched the average with "one" gallery photo from that person. However, there were multiple (from 7 to 28) gallery photos for each test identity in the database (*1*). The experiments did not show the number of single (gallery) photos to which the averages could be matched. In contrast, if the average image is incorporated in identification documents, the identity-verification system must be required to suitably match it to every photo from the same person; otherwise any miss on a photo would translate to a recognition error. Therefore, although the recognition algorithm is commutable, the task of identity verification is more demanding than that of Jenkins and Burton's experiments, and the feasibility of using average images for verifying identity requires further testing. Proof of identity is achieved by comparing an individual's appearance to a photo-identification document, where the appearance is captured by any single facial image in diverse locales and different times. The reliability of the proof depends on how stably the single images can be matched to the corresponding photo-identification document. Hence, in order to evaluate the feasibility of average images on identification documents, a refined experiment should be designed to measure the hit rate for single (gallery) photos, showing what proportion of the single images can be matched to the corresponding average. Moreover, the reliability of the proof also depends on the ability to reject the photos of the impostors according to the averages, which also need to be considered. For the scientific methodology, one can refer to United States government–sponsored evaluations, such as the Face Recognition Vendor Test (*5*), which are the standard test beds for face-recognition technologies.

We acknowledge that image averaging contributes to the face-recognition procedures. However, automatic face recognition is a complex pattern-recognition problem involved with early processing, perceptual coding, and cue-fusion mechanisms (*6*). Although countless solid contributions have been made (*7*), 100% accuracy in automatic face recognition in real-world settings remains an ambitious goal.

School of Information Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China.

*To whom correspondence should be addressed. E-mail: whdeng@bupt.edu.cn

References and Notes

1. R. Jenkins, A. M. Burton, *Science* **319**, 435 (2008).
2. MyHeritage, www.myheritage.com/face-recognition.
3. I. Craw, N. Costen, T. Kato, S. Akamatsu, *IEEE Trans. Pattern Anal.* **21**, 725 (1999).
4. S. Shan, Y. Chang, W. Gao, B. Cao, *Proc. 6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, 314 (Seoul, Korea, 17 to 19 May 2004).
5. Face Recognition Vendor Test, www.frvt.org.
6. P. Sinha, *Nat. Neurosci.* **5**, 1093 (2002).
7. W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfield, *Assoc. Comput. Mach. Comput. Surv.* **35**, 399 (2003).
8. The authors are funded by China Scholarship Council, Natural Science Foundation of China (60675001), and National High-Tech Development Plan of China (2007AA01Z417).

10 March 2008; accepted 15 July 2008
10.1126/science.1157523

Response to Comment on "100% Accuracy in Automatic Face Recognition"

R. Jenkins* and A. M. Burton

Contrary to the suggestion of Deng *et al.*, image registration reduced face-recognition accuracy when divorced from the averaging procedure. Average-to-photo mapping generalizes beyond specific photographs, and averaging either gallery images or probe images can improve the match. The alternative protocol suggested by the authors is unsuitable because it evaluates face-matching algorithms, not face representations, and relies on standard image sets.

We reported that the process of image averaging can dramatically boost automatic face recognition (*1*). Deng *et al.* (*2*) suggest that image registration alone might improve face-recognition performance, and we tested this suggestion. Because the MyHeritage database (*3*) is constantly expanding, we first resubmitted the photographs and average images used in (*1*) to establish a current baseline. Forty-eight of the 500 probe images were identical to images in the online gallery, compared with 41 in (*1*). This increase is consistent with gallery expansion. Of the remaining 452 photographs, 52% were correctly identified, down from 54% in (*1*). The hit rate for the average images was 100%, as before. Five of the average images matched different photos of the correct person, confirming that the average-to-photo mapping generalizes beyond particular snapshots. To address Deng *et al.*'s concern, we next submitted manually registered versions of the source photographs. As Deng *et al.* describe, these were aligned in a standard frontal and upright posture and enclosed by a uniform background. The hit rate for the registered images was 30%. Apparently, registration alone offers the worst of both worlds: It disrupts any informative correspondence in shape between gallery and probe items but does not otherwise stabilize image variability. Registration of the probe images might be less harmful when the gallery images are also registered. In a previous study using a principal components analysis–based image match (*4*), we carried out exactly this transformation. Performance was poor but was nonetheless improved by averaging.

Deng *et al.* (*2*) also express concern that our average images were presented as probes rather than being gallery items. This was a consequence of our chosen methodology. To ensure a stringent test of our averaging technique, we relinquished control over several key aspects of the image match. We used someone else's gallery photographs together with someone else's matching algorithm. Our probe images were collected from the Internet. This approach meant that we were not able to add images to the gallery, but we could still submit images as probes. Because face recognition can be reduced to matching pairs of images, the order of each pair was not our main interest, and we treated matching A to B as equivalent to matching B to A. In previous studies, we have shown that averaging also helps when applied to the gallery images (*4*). Whether identity checks would be better served by an average image stored in an identification document or an average probe generated from the live face is an interesting empirical question. However, it is worth pointing out that averaging probe images specifically finds practical application in forensic face recognition (*5*).

Deng *et al.* point out that an average probe need only match one gallery photo of the target to score a hit. The same is true for the photographic probes, yet these performed comparatively poorly. In practice, an average probe can match very different photos of the target, as our new data confirm. This underscores the major benefit of averaging. Matching pairs of photos is extremely difficult, because both items contain information that is not diagnostic of identity. Matching a photo to an average is helpful because it eliminates nondiagnostic information from one item in the pair. There is no doubt that difficulties can still arise in this situation, but this is partly because the pair still includes a photograph. Our response is therefore not to retreat to matching pairs of photos but rather to investigate ways to eliminate photos from the match altogether. Matching pairs of average images is an obvious route to explore, and we are testing this possibility.

Deng *et al.* recommend the Face Recognition Vendor Test (FRVT) (*6*) as a methodological template. This is unsuitable for several reasons. First, the FRVT evaluations compare performance of different matching algorithms on standard images. Our proposal concerns the representation of the face and is independent of the matching algorithm. Second, the standard databases consist of posed photographs, which grossly underrepresent the variability of ambient face images. Third, reliance on any standard database carries the risk of solving "database recognition" without tackling face recognition. The real world presents different crowds on different days, and systems aspiring to real-world application cannot ignore this inconvenience.

Finally, we agree with Deng *et al.* that early processing and automatic feature extraction are interesting problems, but they are clearly separate from the problem of face recognition. To convince yourself of this, note that it is easy to locate landmarks on a face you cannot recognize and that doing so does not trigger identification.

Department of Psychology, University of Glasgow, Glasgow G12 8QQ, UK.

*To whom correspondence should be addressed. E-mail: rob@psy.gla.ac.uk

References

1. R. Jenkins, A. M. Burton, *Science* **319**, 435 (2008).
2. W. Deng, J. Guo, J. Hu, H. Zhang, *Science* **321**, 912 (2008); www.sciencemag.org/cgi/content/full/321/5891/912c.
3. MyHeritage, www.myheritage.com.
4. A. M. Burton, R. Jenkins, P. J. B. Hancock, D. White, *Cognit. Psychol.* **51**, 256 (2005).
5. V. Bruce, H. Ness, P. J. B. Hancock, C. Newman, J. Rarity, *J. Appl. Psychol.* **87**, 894 (2002).
6. Face Recognition Vendor Test, www.frvt.org.

21 April 2008; accepted 16 July 2008
10.1126/science.1158428

FURTHER READING IS OPTIONAL

(Subsequent pages are just a more detailed explanation of the previous pages)

Final Report: Robust Representations for Face Recognition (R000 23 0437)

Background

Figure 1 shows twenty different images of the same celebrity. These differ on many dimensions, including the nature of the camera, the direction and type of lighting, the age and pose of the person. Despite this huge variability, observers have little difficulty recognising these as the same person. By contrast, we are strikingly poor at recognising unfamiliar faces. (e.g., Bruce et al, 1999, 2001).

The difference between people's abilities with familiar and unfamiliar faces leads to the question of how faces *become* familiar. Since all faces start as unfamiliar, what is it that changes as we learn a face? Previous work suggests a processing shift, whereby the internal features of a face come to dominate recognition, as the person becomes more familiar (Ellis et al, 1979; Young et al, 1985; O'Donnell & Bruce, 2001). Here we explored a different approach. We aimed to test the hypothesis that familiarisation arises through successive refinement of a single stored representation of someone's face.

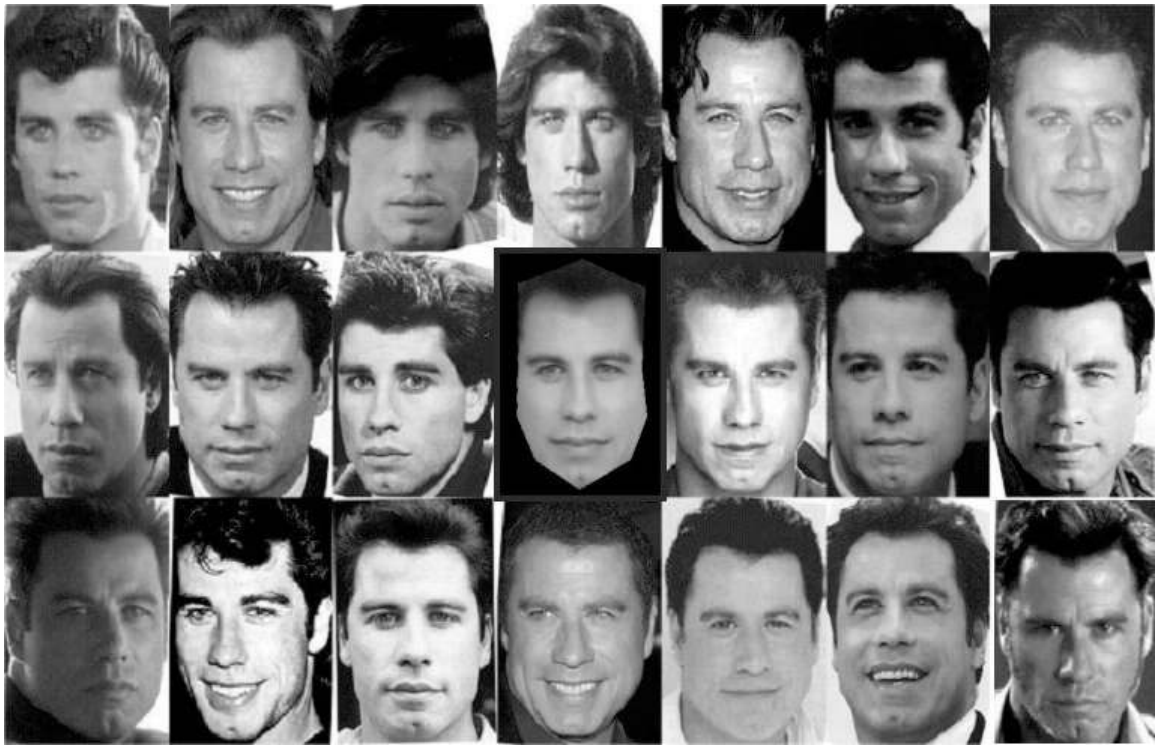


Figure 1: 20 images of a celebrity, the identity-average is shown in the centre.

In pilot work we had developed a representation based on averaging together images of the same person. The resulting average (or prototype) appeared to have some attractive properties. For example, in computer-based recognition, programs based on averages performed better than those storing individual instances of the face separately. Similarly, human observers appeared to recognise the average of a person's face quite well. However, our pilot work was limited for the following reasons:

1. The range of faces tested was severely limited. Construction of the average representations is time-consuming, and we had not been able to develop a sufficiently large database to test these ideas reliably.

2. To construct average face representations, all faces must be morphed to the same common shape (i.e. same outline, features occupying the same part of the image for all faces). This made some of the images poor, while seeming not to damage others. Without a good way of re-introducing shape, the approach seemed limited.

While the focus of the theoretical component of the project was on human face perception, there is a corresponding problem in engineering. There is no automatic system which can tackle the range of variability shown in figure 1, and most systems are developed and tested on images taken under controlled (or at least homogenous) conditions. In a recent review, Zhao et al (2003) state that the problem of face recognition with images captured in variable conditions is so difficult that it should not even be attempted with current levels of understanding. In fact, our observations with the average representation led us to hypothesise that these would be useful for automatic recognition. We therefore built a computer-based face recognition stream into our proposal.

Objectives

The aims in our proposal were:

1. To establish whether a picture-average can, in principle, be used to recognise faces from a wide variation in images.

2. To establish whether this picture-average provides a psychologically plausible account of human recognition of familiar faces.

3. To establish whether incremental update of the picture-average, through new encounters with a face, provides a good account of how faces are learned. Related to this, does the updating provide a good account of the difference between familiar and unfamiliar face processing?

4. To explore the relative contributions of shape and texture to recognising familiar faces.

Each of these has been achieved, as described below. In some aspects, progress was faster than anticipated, and so we have been able to explore some issues further. To the original aims of the project, we added:

5. To explore whether the graphical space occupied by face averages is sufficiently coherent to allow principled manipulations of (i) expression and (ii) identity.

6. To examine hypotheses about the difference between familiar and unfamiliar faces which arise from this work, and specifically to examine memory for faces.

Methods and Results

The format of this section follows the original proposal. However, we should note that one of the most important aspects of the project arose under Strand B, in which we were able to reintroduce shape to the face averages. It quickly became apparent that one method for doing this, using the shape-average, resulted in a powerful representation (Figure 1). The construction of this representation was very productive, and resulted in a paper in the very high impact journal *Cognitive Psychology*.

Image Capture and manipulation

This project relied on obtaining a realistic range of images. We explicitly wanted to avoid images taken under similar lighting, with similar cameras, etc. We therefore planned to capture multiple images of famous people from the internet. This has the advantage that we have no control over the superficial image characteristics, guaranteeing the required variability. It has the further advantage that these images can subsequently be used in recognition experiments with human subjects.

We successfully constructed a database of 1000 full-face images of famous people: 20 different pictures of each of 50 people. The only constraint was that the images should be roughly full-face (to within about 20 degrees). These were all hand-processed to mark key locations (eyes, mouth, outline), and all were morphed to a standard shape. (The resulting faces are called shape-free faces.) Figure 1 shows an example of all twenty faces for one of the celebrities. The top row of figure 7 shows the averages of the shape-free images for each of four celebrities.

Strand A: Prototype face representations: simulation studies

Here we asked whether an average face is capable *in principle* of delivering robust face recognition. In this simulation strand, we compared accuracy for systems whose eigenfaces were derived from *multiple exemplars* or *image averages* of the fifty target faces. The general method for each study was the same. A *model* was built by taking a set of faces (the training set) and generating eigenfaces from these. 50 eigenfaces were generated for all models, regardless of how many faces were used. Each of the training set images was then reconstructed from these 50 eigenfaces. A test set, comprising one novel image of each person was then coded in terms of the same eigenfaces, and a match made against the training set. This match, in 50d space, used a Mahalanobis distance metric, which we have previously found to be essential for PCA to work across images from different sources (Burton et al, 2001). The hit rate for a model reflects the number of times a target image was matched most closely to a training image of the same person.

As proposed, we were able to generate families of simulations by rotating particular faces around conditions, and repeating simulations across different subsets of the database, thus ensuring that results were generalisable. New programs to allow this were developed early in the project.

In **study 1** we compared systems based on 1, 3, 6 or 9 images per person (full details in Burton et al, 2005). Figure 2 shows: (i) that systems are more accurate when they are based on more images of each person; (ii) that systems based on an average representation are better than systems in which all encountered instances of a face are stored separately. This is a very striking finding. Note that exactly the same faces were used in instance- and average-based systems, and yet it is consistently the case that there is an advantage for storing an agglomeration of these rather than storing them individually.

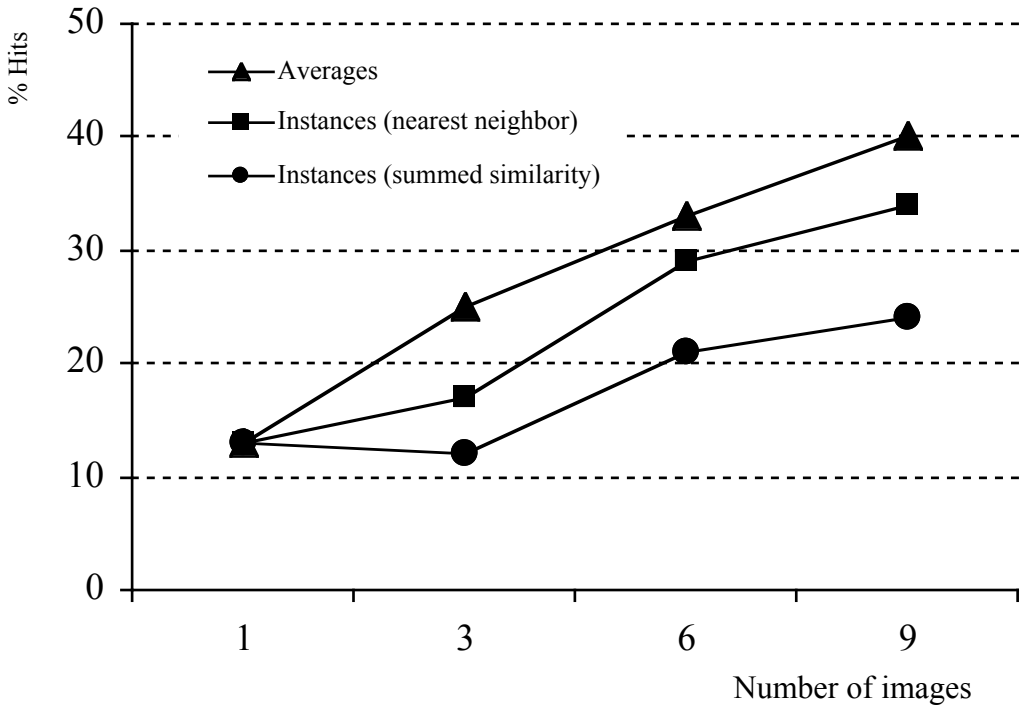


Figure 2: Mean hit rates (%) for systems derived using different numbers of images, for both instance-based and average-based simulations.

In **study 2** we examined the situation in which a single system knows people to varying degrees. All simulations were based on 50 known individuals, but for 10 of these only a single image was stored, whereas for a different 10, an average of three images was stored, and similarly there were ten individuals whose average was based on 6, on 9 and on 19 images. Figure 3 shows performance of such a system (averaged across many simulations, each using a different sub-set of images).

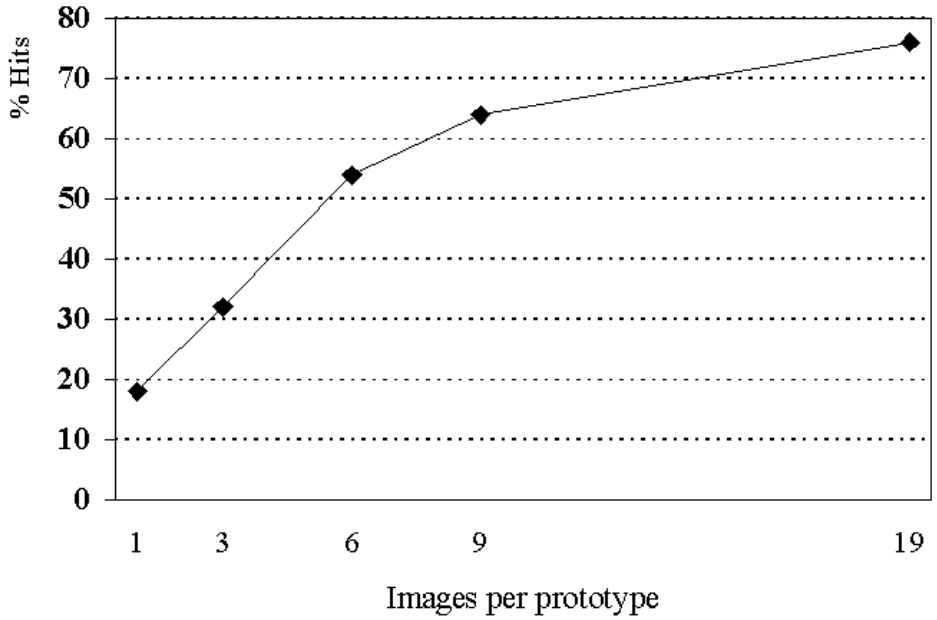


Figure 3 : Hit rate (% recognition) as a function of the number of images constituting each average representation.

The results show that the quality of the average representation continues to improve up to 19 constituent images. Furthermore, the high levels of performance (roughly 75%) are impressive, given the wide range of images used (e.g., figure 1). This level of performance allowed us to write a paper for the IEEE conference on Automatic Face and Gesture Recognition (Jenkins et al, 2006).

In **study 3** we examined how many images are necessary in order to converge on a stable average. Given the range of variability we have in these images, it would be interesting to observe how many are necessary in order to construct a "true" mean. For this study it was necessary to gather more images, and we collected a further 12 images for each of 6 celebrities, giving us a total of 32 images for these celebrities. We then computed a number of averages using the following subsets of each individual's face: sixteen 2-image averages, eight 4-image averages, four 8-image averages, and two 16-image averages. averages. Within an n-image level, no two faces were used twice. For each face we then computed pixel-wise similarity between each average and the reference image (the 32-image average). Figure 4 shows mean pixel difference between all subset averages, and the reference image for one particular identity. For the identity shown (and all others), the four 8-image averages are already highly similar despite the fact that a completely different set of 8 images was used to construct each. It seems then, that this technique converges on a useable mean quickly.

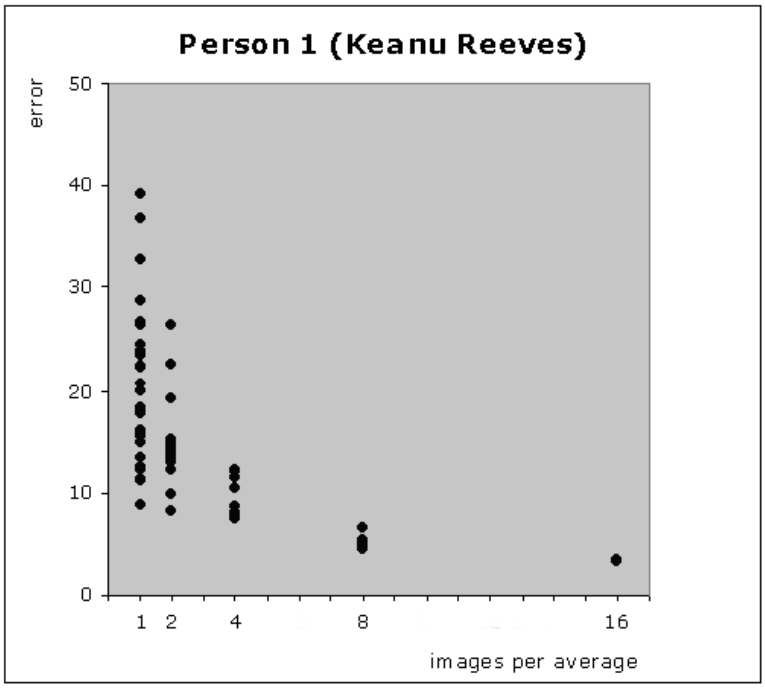


Figure 4 Error (pixel-wise) between reference image, and averages constructed from subsets.

In **study 4**, we asked how robust this representation is against contamination. Does the system begin to fail if the average representation of George Bush is accidentally contaminated by an instance of Tony Blair? (as might happen following a misidentification). We tested this by constructing eigenfaces using different base sets (details in Jenkins et al, 2006). Figure 5 shows average performance of systems in which all known faces are represented as an average of 18 images, or which have been contaminated by mixtures of (a) a random individual, or (b) a similar looking person. The data show accuracy declining relatively slowly, and no catastrophic detriment. This is an attractive property of the representation: errors do not cause outright failures. Furthermore, it is in line with our observations of blending stimuli (see figure 9), in which the average seems to survive contamination even by very different-looking individuals.

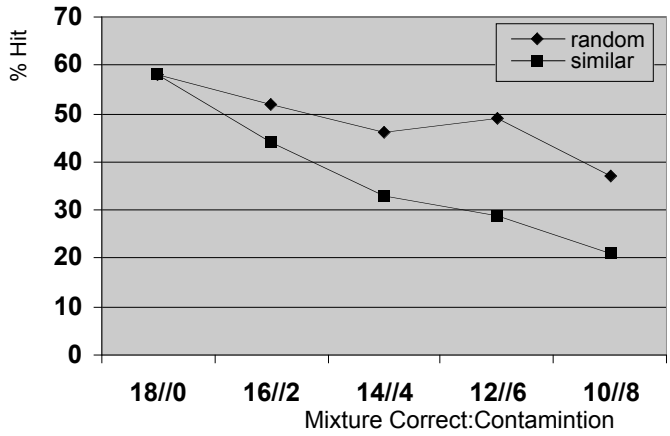


Figure 5: Performance as a function of contaminant nature and proportion

Strand B: Prototype face representations: Human studies.

In this strand we showed human perceivers different versions of the prototypes. We aimed to re-introduce shape into the face averages, though we started the project unclear about how best to do this. In fact, once the software to reintroduce shape had been written, the results were immediately rather clear. If average shape-free faces were morphed back to that person's average shape, a very good representation was formed. We did not expect this. Figure 6 shows the process of extracting the shape and texture for a particular face image. Images such as those to the right are combined to form shape-free averages. We anticipated that these would then form good overall averages if morphed back to the shape derived from a *particular* image (such as that at the bottom of figure 6). In contrast, taking the mean position of each grid-point across a set of images involves taking the average across quite wide ranges in pose, and we anticipated that little individuating information would remain. This turned out to be wrong, and the best representation for human viewing was consistently the average image morphed to the average shape, a representation which we refer to as an *identity-average*. Figure 7 shows some examples, and figure 8 shows the average faces, prepared in this way, for our entire database of 50 celebrities.

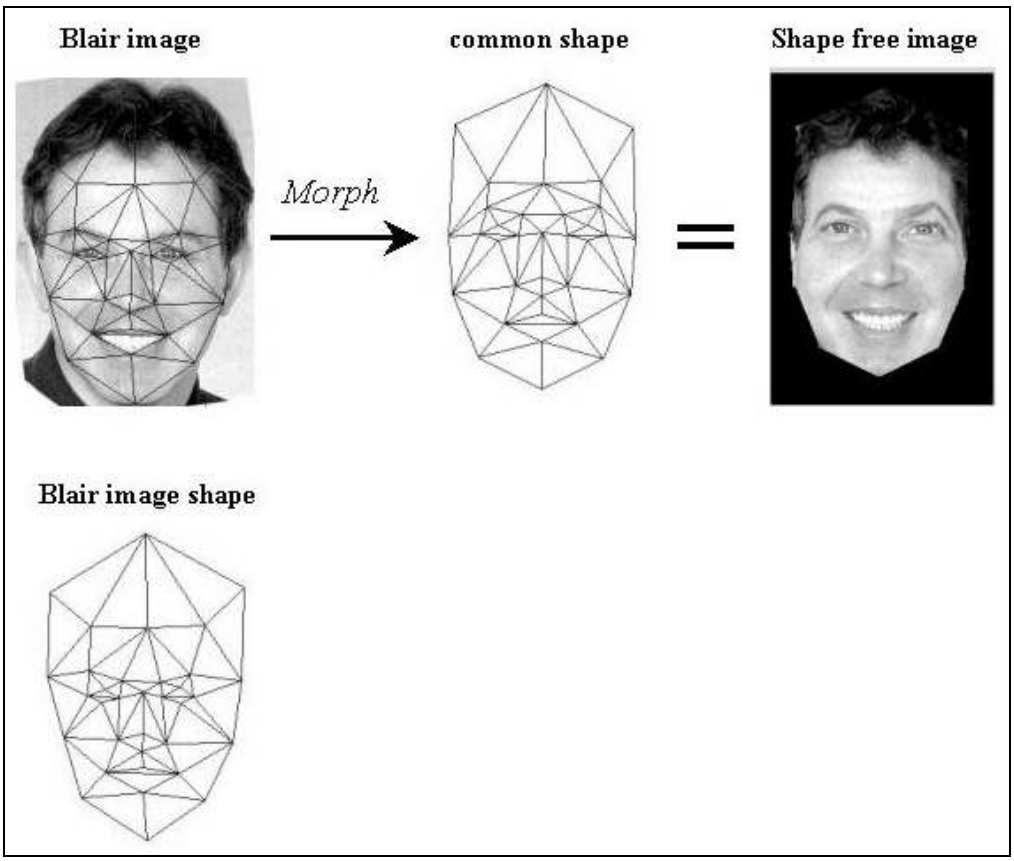


Figure 6: decomposition of a face image into shape and texture components

Figure 7: Averages of four celebrity images. The top row shows the averages of shape-free images. The bottom row shows the results of morphing these to each individual's average shape.

Figure 8: Averages for 50 celebrities (20 images of each) constructed to include their shape

We will summarise experiments performed on these types of images in the remainder of this section.

Studies 5-8 all had the same form. Using verification procedure, subjects saw a name, followed by a face. The task was to indicate (by button press) whether the face corresponded to name. In these experiments we manipulated the form of the face: it may be a photograph, a shape-free average, an identity-average, etc, according to the particular experiment. Further details of the procedure are reported in Burton et al, (2005).

In **study 5** we compared name-verification accuracy and speed for (i) shape-free average faces; (ii) average faces morphed to the shape of a specific exemplar of that person; (iii) average faces morphed to that person's average shape (identity-averages). Accuracy in these conditions was 65%, 84% and 86% respectively. Reaction times showed a similar pattern (821ms, 746ms, 741ms respectively). These data clearly demonstrate that the introduction of shape improves the representation over the shape-free version, though the advantage for the identity-average is only small by comparison to an image morphed to a shape-instance. In **study 6**, we therefore manipulated the instance/average dimension independently for both the face-texture and shape, giving a factorial design. Accuracy data showed that the identity average was recognised best (16% errors) with a photograph (instance face, instance shape) being recognised less well (21%), with similar accuracies for the instance-texture/shape-average (21%) and average-texture/shape-instance (23%) errors. This is the first example of an identity-average being recognised better than a real image, and we were to repeat this pattern in later experiments. In **study 7** we replicated the two key conditions of study 6, but with a larger stimulus set (in order to satisfy reviewers of Burton et al, 2005). Comparing identity-averages, to average-faces morphed to a shape-instance, we found an overall advantage for identity averages in both accuracy (89% vs 80%) and RT (700ms vs 720ms).

In **study 8** we found the same advantage for identity-averages, even when all images were stretched to twice their height, a result consistent with data presented by Hole et al (2002).

In a further test of identity averages, we used a different procedure in **study 9**. We presented subjects with printed images, and asked them to identify the face, either by name or other individuating information. Under no time-pressure, subjects showed an advantage for recognising identity averages over real photographs (81% vs 77%). This is a small effect, but is exciting because recognition of familiar faces is normally so good that it is hard to find stimuli which give even higher levels of performance. Here we have an example of something akin to a super-stimulus, and the effect cannot be attributed to image characteristics, since the stimuli were rotated around conditions across the experiment (i.e. Bill Clinton appeared equally often as a photo and an average). Our current hypothesis about this effect is that it reflects the "smoothing out" effect of averages. For example, it is possible to have a photo which is not a particularly good likeness of Bill Clinton, whereas it is impossible for an average to have this characteristic. This might result in a small effect of the type observed here.

As outlined in the proposal, we also explored the use of averages in priming studies. Results of these studies have not proved as exciting as the studies above, so we summarise them only briefly here. In **study 10** subjects saw a set of faces, half of which were photos, and half identity-averages. An unexpected test phase followed, in which subjects saw new photos of the same celebrities, intermixed with unknown faces. They made speeded familiarity decisions to these faces. Results showed a slight advantage in RTs for faces primed by an identity-average (792ms vs 806ms), though this did not reach significance (despite a powerful design: 40 subjects, within subjects, 24 test items per condition). In **study 11** we repeated this with a different subset of stimuli, but again found no reliable difference between conditions. In **study 12** we adopted a short-term self-priming approach (cf Calder & Young, 1996). Consistent with the literature, we observed large and reliable self-priming effects (60ms), but these were unmodulated by presentation of either a face photo or an identity-average. In **study 13** we repeated the previous experiment, but this time used a short SOA (50ms as opposed to 250ms), a manipulation which has been shown to increase priming in this situation. We found reliable priming, but no modulation by prime-type. Finally, in **studies 14 and 15** we repeated experiments 12 and 13, using a different set of exemplars for the photograph condition. Once again, no effects of prime type were observed, despite large and reliable priming effects. In one sense the results are encouraging, because across all these five experiments identity-averages always produced the same effects as real photographs, and so acted as good representations. However, we were not able to demonstrate the more exciting result of an advantage for averages using priming.

Strand C: Face learning

As proposed, we have used a variety of techniques to study face learning (see Strand A). In one approach, we have asked how robust is an average when contaminated by instances of the wrong person. Figure 9 shows an example of mixed-identity averaging. In each case, the averages are formed from 20 images. At the top left, all are images of Marilyn Monroe. At each step, one MM image is dropped, and an Elvis image added. Although the two people are highly dissimilar, the transformation is smooth, and we have already shown for automatic systems (**study 4**) that the degradation in performance is smooth. With humans (**study 16**) we used mixed-identity stimuli (i.e. from the mid-point of the continuum), and encouraged subjects to "see" one of the people. In a subsequent test phase, we observed priming from both "seen" and "unseen" identities in the image, though this is rather small. We will follow up this research in the future.

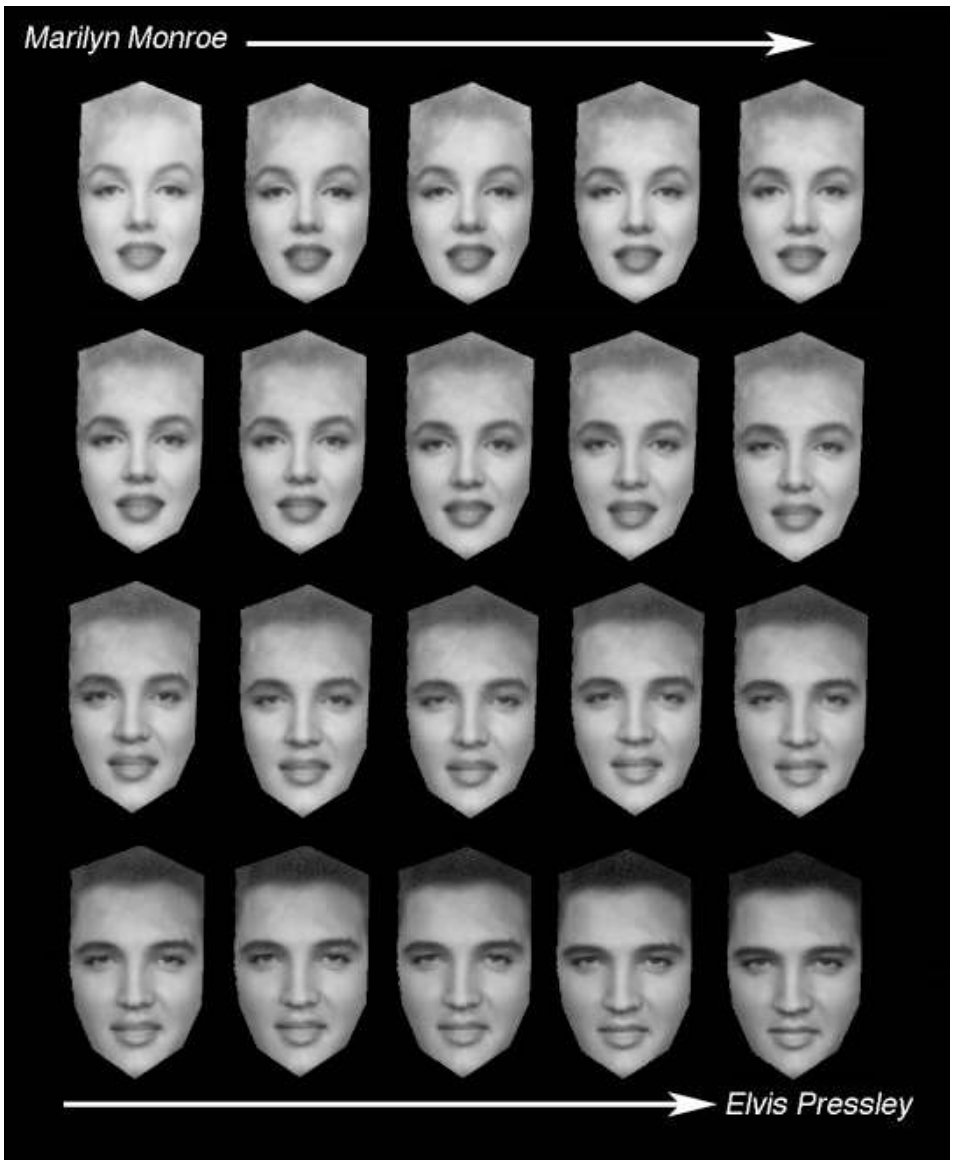


Figure 9: Each face is the average of 20 images. Top left are all Marilyn. Bottom right all Elvis and intermediate averages are mixtures.

Further investigations have been triggered by an observation by our RA, David White. When compiling the database he noticed that it is difficult to remember whether one has already gathered a *particular image* of a celebrity, and that this seemed not to be the case for unfamiliar faces. This is consistent with our hypothesis that unfamiliar faces are processed as images (i.e. one's representation preserves surface features) whereas familiar faces are processed for *gist* (i.e. do not preserve surface features), see Hancock et al (2000). To test this possibility, we ran a series of experiments with the same design. Subjects were shown a sequence of 7 faces, all of the same identity. They were then shown an eighth, and asked whether they saw *this particular image* of the face. Figure 10 shows the results from **study 17**, using this design. In summary, only unfamiliar faces show a primacy effect for this task. It is extremely unusual (possibly unique) to observe an advantage for unfamiliar faces in any task, and the use of a task relying on memory for specific images seems to provide powerful evidence that unfamiliar faces are processed in qualitatively different ways from familiar faces. We pursued this effect using inverted faces, verbal material, and photographs of familiar and unfamiliar buildings (**studies 18-20**, respectively). Only faces showed this effect. A paper on these unusual findings was presented at EPS in April 2006, and we will follow this up with further experiments, prior to submitting a paper for publication.

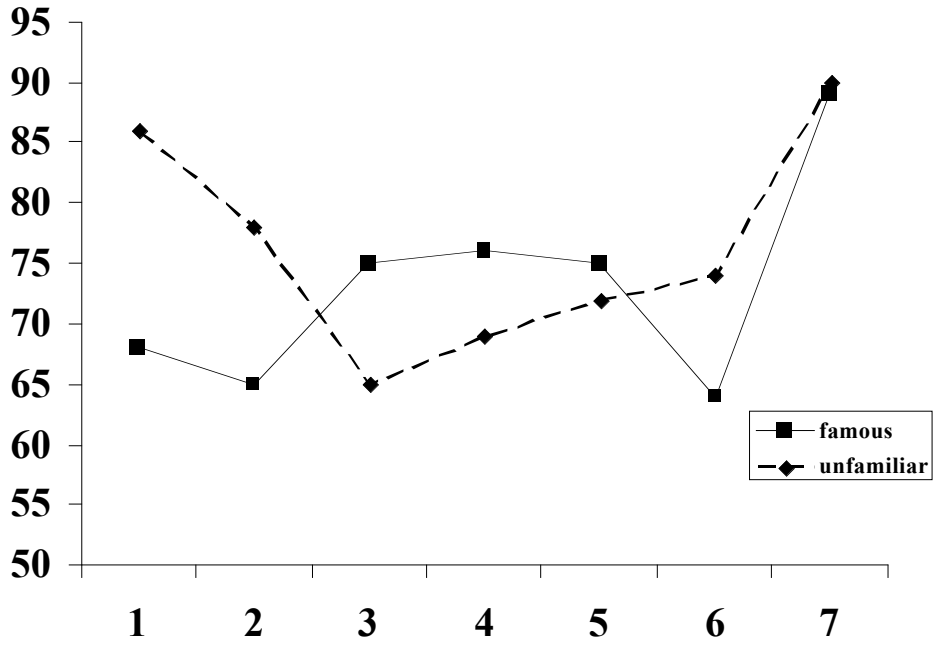


Figure 10. Recognition accuracy as a function of serial position

Further studies

Progress on some of the planned work was quicker than expected, and we have therefore been able to pursue a number of interesting lines of research. We have begun to apply the averaging principle to other dimensions of face perception besides personal identity. This extends our general approach of morphing disparate images into a common linear space for processing, and then morphing back out of that linear space once processing is completed. This approach is extremely powerful, as it allows visual information from entirely separate sources to be combined using simple image algebra. Applying averaging to the separate streams of images before they are combined renders them compatible in terms of lighting dimensions, as within each stream lighting effects will tend to cancel out as more images contribute to the average. In practice, this means that the streams to be combined share neutral lighting. Figure 11 illustrates how we have used this technique to extract emotional expression information from one source and transfer it to faces from a completely separate source. Data from human observers shows that identity and expression can be easily read from the resulting images (see Figure 12). The finding that both signals can survive recombination suggests that our approach may shed light on the long-standing issue of independent processing of identity and expression in face perception. This work was presented at the London meeting of EPS 2006.

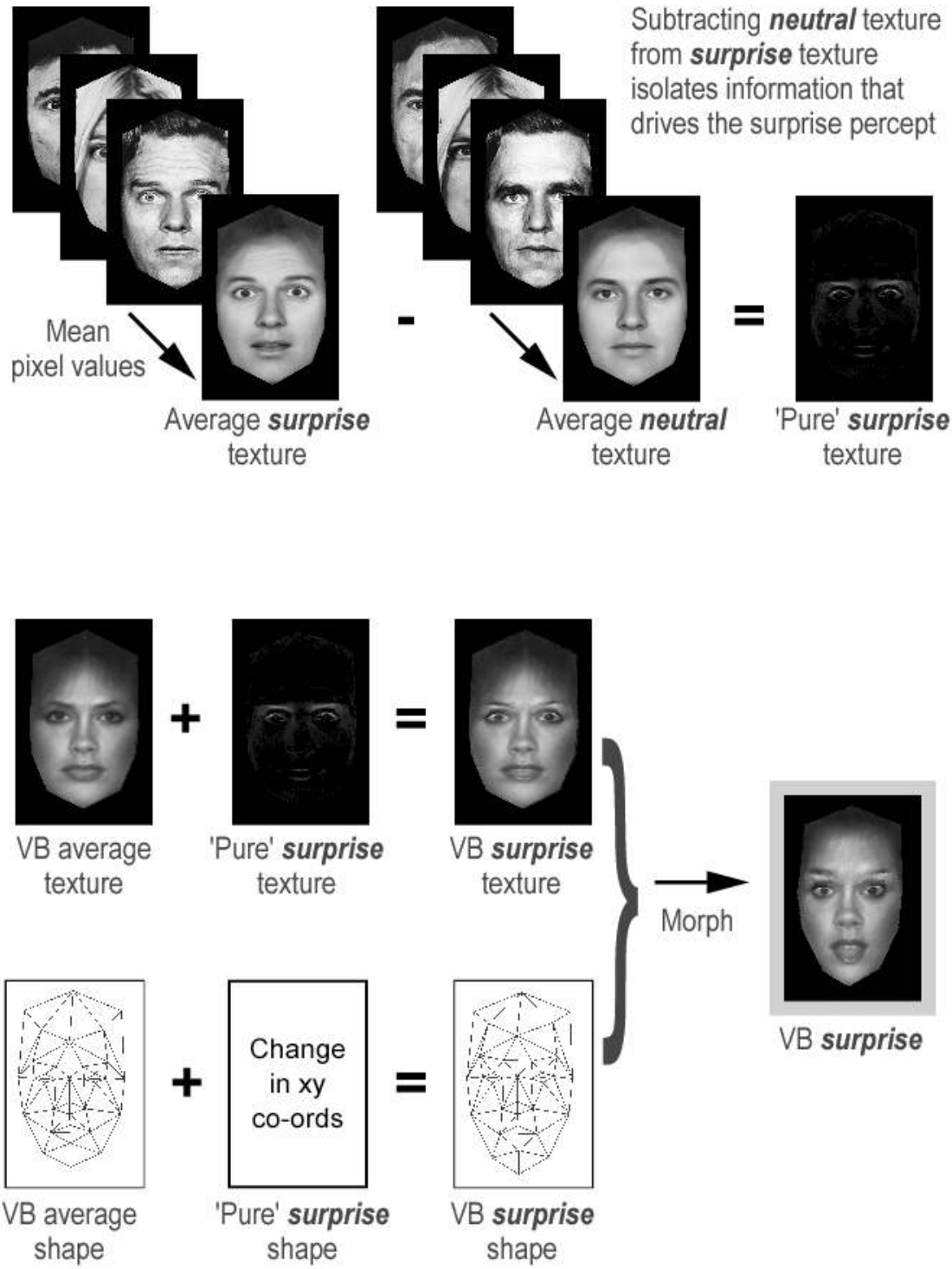


Figure 11: Pasting a surprise onto a neutral face

Figure 12: Six expressions, derived from the Ekman and Friesen set, but applied to celebrities.

We have recently extended our approach to examine further the issue of information underlying perception of specific facial identities. The technique involves comparison of a given *identity average* against an *overall average* face computed across all identities in our database (the *population mean*; see Figure 13). The subtraction *Victoria Beckham minus population mean* leaves a difference image that shows how Victoria Beckham's face differs from the average face. In other words, if this individuating information is *added* to the population mean, the result is Victoria Beckham's face. Our new observation is that if the same information is *subtracted* from the population mean, the result is a face with contrasting appearance – Victoria Beckham's *antiface* (see Figure 13). This antiface has some interesting characteristics. First, it looks like a plausible photographic face. It was not obvious in advance that this would be the case. Second, psychologically relevant dimensions such as sex and emotional expression are reversed by this process (female becomes male; sullen becomes cheery), even though these dimensions are not explicitly coded at any stage. In addition, all aspects of the physical appearance of the face take on the opposite valence, so that dark complexion becomes light complexion, upturned nose becomes downturned nose, etc. One very useful property of these antifaces is that they exactly match their positive counterparts in terms of inter-item similarity. That is, anti-Victoria Beckham and anti-Margaret Thatcher are physically exactly as similar to each other as are Victoria Beckham and Margaret Thatcher. This is an extremely desirable feature of the image set, as it will allow us for the first time to compare perception of familiar and unfamiliar faces that are matched in terms of these physical attributes.

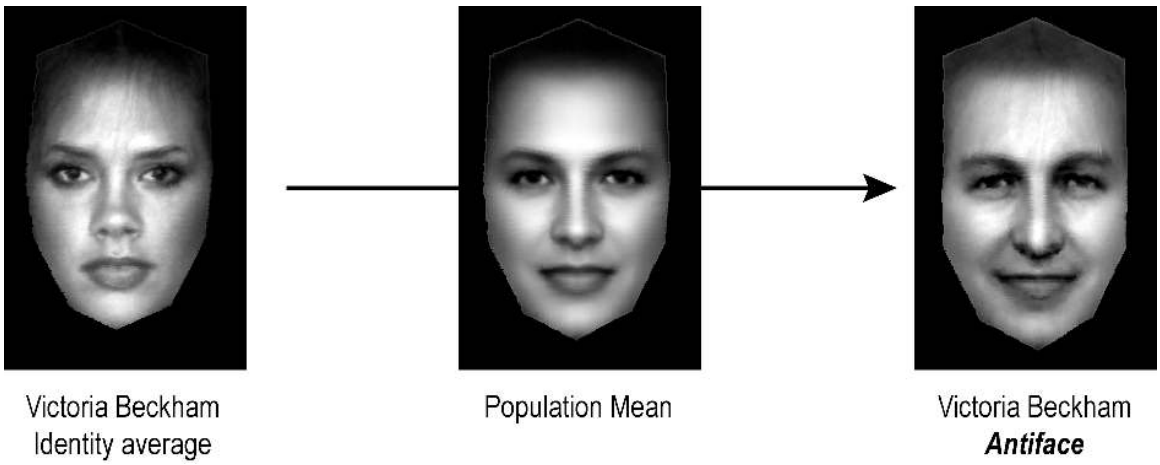


Figure 13: Constructing antifaces

One way of visualizing the relationship between faces and their corresponding antifaces is to consider each face as a point in a high-dimensional space. By definition, the population mean occupies the origin of this space. Similarly the average male and female faces are located at different positions. Figure 14 shows a schematic diagram of this face space. Projecting Paul McCartney's face through the population mean generates his antiface as described above. Note that this antiface is in the female region of the space, and has a feminine appearance. As well as generating this opposite-sex antiface by projecting the original face through the population mean, it is possible to generate a same-sex antiface by projecting the original face through the same-sex mean (the male average in this case). Paul McCartney's same-sex antiface is inside the male region, and so has a male appearance, but it is in the opposite quadrant of male region. In fact, it occupies the same point in the male region that the opposite-sex antiface occupies in the female region. This means that the two antifaces look alike in all respects, except that one looks male and the other looks female. Note that an original face and its anti-face are equally legitimate faces in this space. The only difference in this example is that the former is familiar, whereas the latter is not. However, their equivalence permits us to do to an antiface anything that we might do to its original. For example, we could take Paul McCartney's opposite-sex antiface as our starting point, and project this through its same-sex mean (the average female face) to arrive on the opposite side of the female region. In fact, this is the same point in the female region that Paul McCartney occupies in the male region. The result should therefore be a face that looks like Paul McCartney's, but is female. As can be seen from Figure 14, this is indeed the result.

We have begun to use stimuli such as these in studies on adaptation effects (see Leopold et al, 2001). Our initial results confirm that perception of the overall average face can be biased by adaptation. For example, the average face is seen as Paul McCartney following exposure to Paul McCartney's antiface. Yet the same average face is seen as Clint Eastwood following exposure to Clint Eastwood's antiface. These findings provide the first demonstration of adaptation effects involving familiar faces. Replication of these effects in full-scale experiments will provide new ways of tapping the structure of our internal face representations.

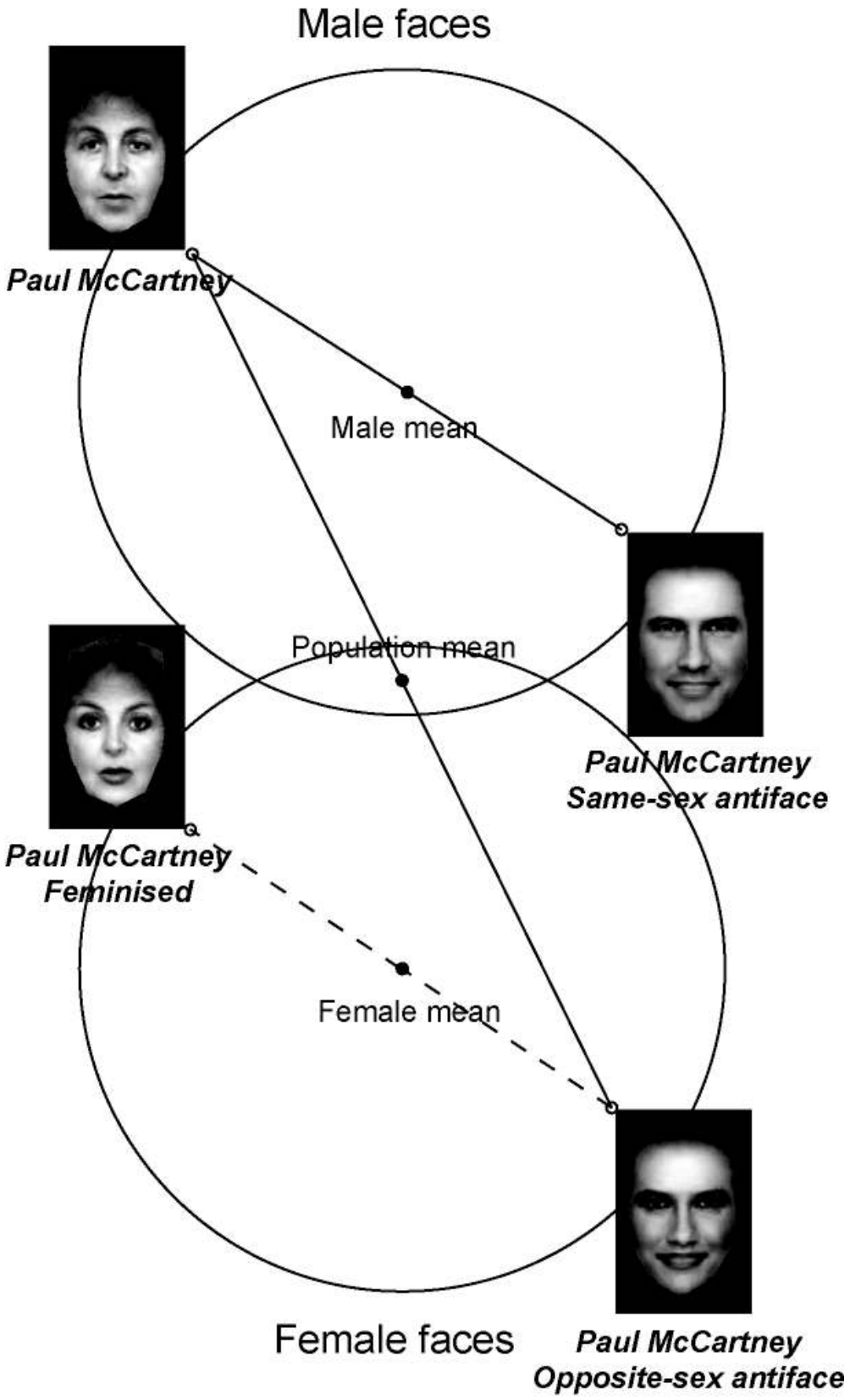


Figure 14: Exploring the space of faces

Our final development is the observation that the space developed here can be used in relation to the identity average for a *particular* face. Figure 15 shows a schematic of the Tony Blair subspace in which individual Blair images form points in the space, and the Blair average defines the origin of the subspace. Projecting any given Blair image through his identity average results in a new image of Blair that has contrasting *image* characteristics. For example, anti-images exhibit the opposite pattern of lighting, facial expression and pose (see Figure 15). We have not yet conducted any experiments based on this observation, but it is clearly very much in tune with our original proposal, and offers another direction in which these ideas developed in this grant can be extended.

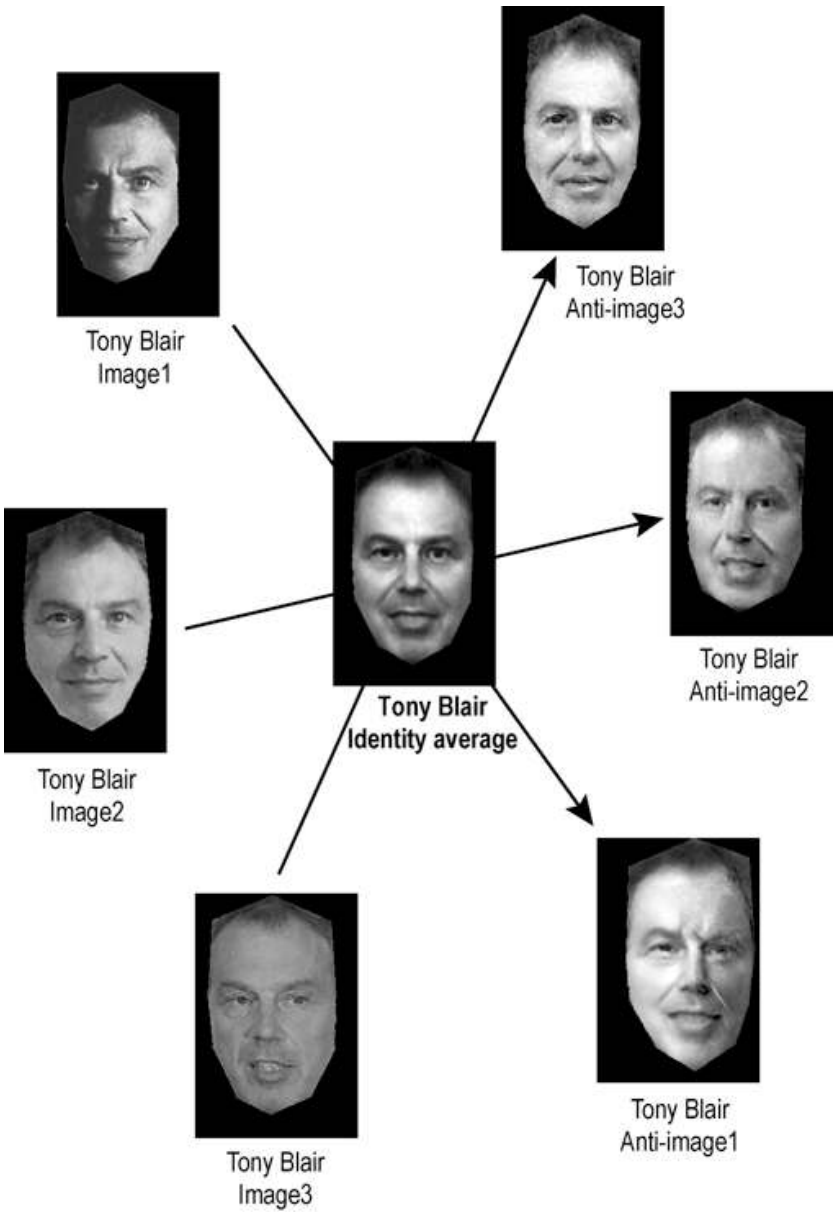


Figure 15: Blair images and some new Blair images

Activities and outputs

Two significant papers have already been published, and these accompany the report:

Burton, A.M., Jenkins, R., Hancock, P.J.B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51*, 256-284.

Jenkins, R., Burton, A.M & White, D. (2006). Face recognition from unconstrained images: progress with prototypes. *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition.* IEEE, 25-30.

The first is a high-impact psychology journal, and the second is the primary international outlet for engineering-based research in face recognition. We have therefore targeted two different communities with different aspects of this work, and the papers contain full theoretical developments impossible to detail in this brief report. Further papers are planned on the basis of work described above.

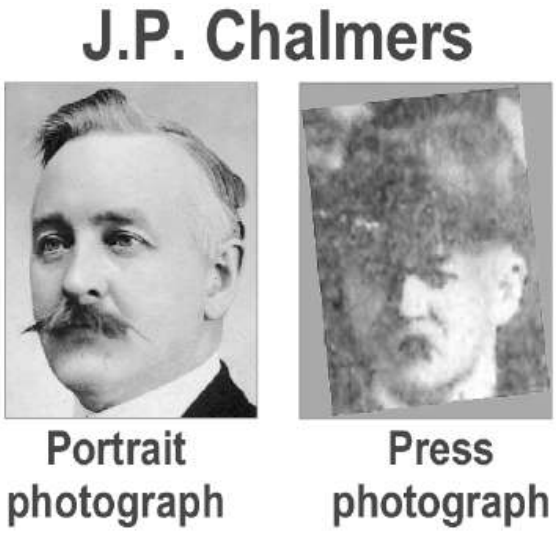
In addition, the following presentations have been made on work contained here.
Burton: EPS Montreal (July 2005) and Birmingham (April 2006); British Neuropsychological Society, London (November 2004); Dana seminar (London Science Museum, November 2004); departmental seminars at Sheffield, Lancaster, Brunel, Essex, Royal Holloway, University of New South Wales, Macquarie, University of Western Australia.
Jenkins: EPS London (January 2004 and 2006); British Academy, London (March 2005); BPS Cognitive Section, Leeds (September 2005); BAAS, Dublin (September 2005); departmental seminars at MRC Cognition & Brain Sciences Unit, Cambridge; University of Westminster, London.

Impacts

Our major papers have been published only recently, though have attracted considerable interest from the community. We have been asked for copies of our database by four groups (two UK, one in Spain, one in USA). We have also been asked for the software written on this project for constructing average faces (Universities of Taiwan and Kyoto). We have discussed with various bodies the possibility of developing the theory developed here into a commercially-viable automatic face recognition system.

We have also made a small contribution to a biographer who contacted us about the cinematographer J.P. Chalmers. Chalmers grew up in Orkney, later leaving for Hollywood. The author's problem was that there are few surviving photos of him. She had established that in his early adulthood, Chalmers worked at the Orkney Herald newspaper, and the Herald took annual staff photographs, making it possible that he appeared in one of those. Figure 16a shows images known to be Chalmers, while 16b and 16c show staff photos from the relevant years. This case provides a real-life example of the kind of unfamiliar face matching task that people find difficult. However, using techniques developed here, we were able to establish with relatively high probability that Chalmers is person 29 in Figure 16c. Using our PCA technique, person 29 was shown to be the best match for *both* images in 16a. For comparison, we showed the same images to visitors to the Glasgow Science Centre. There was very large variation in choice, though person 29 was picked most frequently as a match for the good quality photo in 16a. However, human subjects were utterly unable to match the poor photo in 16a. Work on the Chalmers case was presented at the *Seeing Faces in the Brain* symposium at the London meeting of EPS in 2004. It was also presented as an invited talk at the British Academy in 2005 and at the Cognitive Section of the BPS, 2005.

a



b

c

Figure 16: Two pictures known to be Chalmers (a) and two staff photos

Future Research Priorities

Many lines of enquiry remain open. Our major questions include:

1. Can the automatic face recognition results gained so far be operationalised into a genuinely useful automatic tool? Since automatic systems currently avoid realistic levels of variation such as those we have used, this is a promising issue. We will seek further funding to explore this.

2. Can this proposal be used to generate further testable predictions about face learning? We wish to develop this work further and ask whether the theory can be used to understand new questions in face learning such as: how many (what range of) faces is needed to become familiar? can one speed face learning by selecting examples? can effects such as the other-race effect, distinctiveness and relations between expression, sex and identity be understood with a common averaging mechanism? These questions as these are broader than we have been able to address so far, and we are likely to seek further funding to examine them.

Word Count: 4960

Annex: References

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.

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., Jenkins, R., Hancock, P.J.B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology, 51,* 256-284.

Calder, A.J. and Young, A.W. (1996) Self priming: a short-term benefit of repetition. *Quarterly Journal of Experimental Psychology, 49A*, 845-861.

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.

Hancock, P.J.B., Bruce, V. & Burton, A.M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Science, 4(9)*, 330-337.

Hole, B G.J., George, P.A., Eaves, K., and Razek, A. Effects of Geometric Distortions on Face Recognition Performance. *Perception, 31*, 1221-1240.

Jenkins, R., Burton, A.M & White, D. (2006). Face Recognition from Unconstrained Images: Progress with Prototypes. *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition.* IEEE, 25-30.

Leopold, D. A., O'Toole, A. J., Vetter, T. & Blanz, V. 2001 Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*, 89–94.

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

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.

Zhao, W., Chellappa, R., Rosenfeld, A. Phillips, P.J. (2003). Face Recognition: A Literature Survey. *ACM Computing Surveys,* 399-458