PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY B

Phil. Trans. R. Soc. B (2011) **366**, 1671–1683
doi:10.1098/rstb.2010.0379

Review

Stable face representations

Rob Jenkins* and A. Mike Burton

Department of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QQ, UK

Photographs are often used to establish the identity of an individual or to verify that they are who they claim to be. Yet, recent research shows that it is surprisingly difficult to match a photo to a face. Neither humans nor machines can perform this task reliably. Although human perceivers are good at matching *familiar* faces, performance with *unfamiliar* faces is strikingly poor. The situation is no better for automatic face recognition systems. In practical settings, automatic systems have been consistently disappointing. In this review, we suggest that failure to distinguish between familiar and unfamiliar face processing has led to unrealistic expectations about face identification in applied settings. We also argue that a photograph is not necessarily a reliable indicator of facial appearance, and develop our proposal that summary statistics can provide more stable face representations. In particular, we show that image averaging stabilizes facial appearance by diluting aspects of the image that vary between snapshots of the same person. We review evidence that the resulting images can outperform photographs in both behavioural experiments and computer simulations, and outline promising directions for future research.

Keywords: face perception; face recognition; identification; security; familiarity; biometrics

1. BACKGROUND

The modern psychological study of face recognition has its roots in the problem of eyewitness testimony. In the 1970s, it became clear that witnesses to a crime could very often be mistaken when subsequently asked to remember someone involved. It is now well-understood that witnesses can make honest errors when trying to recall a face to use in a description or a reconstruction (photofit, e-fit, etc.). They may also make errors in recognition memory, for example, when asked to recognize a criminal in a police mugshot or an identity parade. This observation led to an explosion of research demonstrating the very many factors that can influence accuracy [1–5].

Research in face recognition for legal purposes has been inspired by legal processes, but has also informed them. For example, in the English legal system, jurors are routinely informed that eyewitnesses can be mistaken in good faith, and that their confidence in face identification need not indicate its accuracy. Despite this, the problem persists, and face-identification errors continue to cause problems for the legal system. The Innocence Project (http://tinyurl.com/8dshn) is a US organization aiming to exonerate wrongfully convicted people using DNA evidence not available at the time of trial. At the time of writing, over 240 people have had convictions overturned, and critically, in 73 per cent of these cases, incorrect eyewitness testimony has been the key evidence.

The early applied emphasis on eyewitness testimony led to a predominance of memory-based theorizing in psychological work on face recognition. In the 1970s and 1980s considerable progress was made in understanding lexical aspects of language, i.e. word recognition and retrieval [6–8]. The style of theoretical reasoning which had proved useful in word recognition, was also being employed in object recognition research [8–10], and this influenced early theoretical models of face recognition. The canonical model, by Bruce & Young [11], followed early developments along similar lines [12,13]. However, what all these models had in common (and where they differed from analogous models of object recognition) was the observation that faces served multiple signalling purposes. Viewers can clearly derive identity information from a face (if they know the person), but they can also derive information about emotional state, facial speech, focus of attention and so on. In early research on face processing, the derivation of identity was the most popular topic for research. In modern times, research has a much broader focus, and for a great deal of current research, identity is irrelevant (see this issue).

In parallel to early psychological theorizing about face recognition, technical developments in image processing made possible the study of automatic, computer-based face recognition. Once again, early studies focused almost exclusively on derivation of identity [14–16], while more recently, this focus has broadened. However, unlike the study of human face processing, the automatic case has continued to be dominated by studies of identity. We will discuss some recent approaches later, but the reason is relatively clear in a security climate in which surveillance and restriction of access are important political imperatives.

In the remainder of this paper, we will review current approaches to understanding how viewers can

* Author for correspondence (rob.jenkins@glasgow.ac.uk).

One contribution of 10 to a Theme Issue ‘Face perception: social, neuropsychological and comparative perspectives’.

derive a person's identity from their face. However, before doing so, it is worth noting that, just as in 1970s, technological and social factors are influential in determining the topics for research. In the present day, many Western societies are preoccupied with proof of identity. It is increasingly common to be asked to prove one's identity by production of photo-ID. This is clearly an identification task ('is the person carrying the correct ID?') but does not rely on memory—at least to the extent that unreliable memory is a well-established problem for eyewitness testimony. As we will describe below, when the viewer is unfamiliar with the person presenting such ID, this turns out to be a surprisingly difficult task. This issue of face *matching*, as opposed to face memory, is also raised by changing forensic technology. The use of closed-circuit television (CCTV) is now commonplace and, at first sight, it might have been predicted that this would solve many of the problems of eyewitness memory. After all, with a complete record of the event, including photos of any perpetrators, the problem of fallible memory does not arise. However, as we will describe below, the problem of *matching* CCTV images to suspects has proved to be much harder than originally anticipated.

2. MATCHING UNFAMILIAR FACES: MACHINE PERFORMANCE

Photo-ID documents continue to be central to national security policies. A number of European countries have introduced a national ID card that includes a photograph of the holder, and a similar scheme was launched in the UK in 2009 (although it is under review at the time of writing). UK passports already include a digital copy of the photograph of the bearer. The intention is that these image files will be machine read and compared with the face of the traveller (e.g. the SmartGate deployed by Australian Customs at Sydney airport). For several reasons, however, the advent of machine systems does not represent a solution to face recognition. Although performance of machine systems on benchmark tests (e.g. Face Recognition Vendor Test) has improved in recent years, today's best systems are far from infallible. This is true even under highly restricted testing conditions, based on tightly controlled, high resolution images and cooperative subjects. Under these optimal conditions, accuracy levels as high as 99 per cent have been achieved [17,18]. The problem is that when conditions are not so favourable (as in border control or surveillance settings), or cooperation is poor (for example, when someone is trying to conceal his or her identity), performance plummets. Accordingly, real world deployments in the USA (http://tinyurl.com/358a4jf, http://tinyurl.com/m6ml), the UK (http://tinyurl.com/2u9epwg, http://tinyurl.com/ccauop) and Australia (http://tinyurl.com/2utuemy), have drawn considerable popular criticism over high error rates. In at least one case (http://tinyurl.com/358a4jf), the poor performance reportedly led to the eventual withdrawal of the scheme without a single recognition hit having been recorded.

It is perhaps easy to see how this situation could come about. Security and surveillance systems are typically concerned to minimize 'miss' errors, for example, to ensure that fraudulent documents are not accepted as legitimate. Unfortunately, this means that they tend to generate high numbers of 'false alarm' errors, in which legitimate matches are challenged. The task of rechecking all the queried cases then falls to human operators, who make the final decision. We return to the issue of human performance limits in §3. For now, the practical problem is the reintroduction of the processing bottleneck that automation is intended to ease. According to one recent UK report (http://tinyurl.com/ccauop), a high false alarm rate in the automatic face recognition system at Manchester Airport was causing unsustainable delays. The response was allegedly to recalibrate the system so that it would admit even passengers with a very poor resemblance to their passport photographs, effectively switching the machine off.

How are we to account for the discrepancy between the rather impressive performance of automatic face recognition systems on benchmark tests [17,18] and their unusable performance in the real world? Evidently, one of these situations does not capture the applied problem. And it is not the real world. Benchmark tests can certainly be useful, for example, when comparing performance of different face matching algorithms on a standard image set. But performance on a benchmark test does not straightforwardly translate to performance on the modelled task. There are a number of reasons for this disparity. For example, in databases of posed images, taken under similar conditions with similar cameras, *within*-person variability in appearance will normally be smaller than in real world samples. At the same time, benchmark databases might over-represent diversity *between* individuals, as their limited size reduces the likelihood of similar pairs. In addition, reliance on any standard database carries the risk that developers might solve 'database recognition' without tackling face recognition. The real world presents different crowds on different days. Systems aspiring to real world application must confront this practical problem.

Although such database construction issues will tend to inflate estimates of machine performance, there is a more fundamental point that has been largely overlooked: to some extent, the disappointing performance of automatic face recognition systems may reflect unrealistic ambitions on the part of developers. By requiring systems to match pairs of photographs, they are setting a problem that human observers find extremely difficult. We suggest that a major attraction of using facial appearance to establish identity is that we accept it can be done in principle. In fact, we experience practical success every day because the system that has solved it is the human brain. The proliferation of 'biologically inspired' approaches to automatic face recognition reflects the willingness of computer engineers to model the brain's success. Yet, psychological studies have shown that human expertise in face identification is much more narrow than is often assumed. Moreover, the process that most automatic systems attempt to model lies outside this narrow expertise. From this perspective, disappointment in machine systems is inevitable, as they model a process that fails. Human limitations in face identification are not widely appreciated even within cognitive psychology, and seldom penetrate cognate fields in engineering and law. In §3, we offer an overview of the most pertinent limitations. For this purpose, we focus specifically on evidence from face matching tasks, as these directly address a problem that is common to security and forensic applications.

3. MATCHING UNFAMILIAR FACES: HUMAN PERFORMANCE

Psychological research has shown that it is surprisingly difficult to match a face to an image. This routine task, which is performed hundreds of times every day by passport officials, security personnel and police officers, turns out to be highly error-prone. In one of the early demonstrations of this, Kemp *et al.* [19] carried out a field test to establish the level of fraud protection afforded by the inclusion of ID photos on credit cards. Supermarket check-out staff were recruited to validate the photo-credit cards by deciding whether or not the photograph was of the person presenting the card. Even though the staff were aware that they were taking part in a study concerning the utility of photo-credit cards, they performed surprisingly poorly. About half of the fraudulent cards were accepted, and about 10 per cent of the valid cards were falsely rejected. More recent laboratory-based studies have replicated this basic finding. Megreya & Burton [20] reported an error rate of 17 per cent for matching recent photos to live faces. Davis & Valentine [21] asked participants to match live persons to CCTV clips. As with the preceding studies, observers were highly error-prone on this task, even when the CCTV footage showed high quality, recent, close-up sequences.

The problem persists when viewers are asked to compare static photographs. In a pioneering demonstration of this, Bruce *et al.* [22,23] devised a task designed to model a best-case scenario for identifying images captured on security video. Participants were shown an array of 10 faces along with a target face. Viewers were asked, for each array, whether or not the target person was present among the 10 candidates, and if so, to point out the match. In the original experiments, the target was present on half of the trials, and absent on the other half. One way to think of the array is as a photographic version of police line-up. As with real line-ups, all the faces fit the same general description (they were all clean-shaven young men with short hair). Participants performed surprisingly badly on this task, with error rates of 30 per cent in both target-present and target-absent conditions.

This poor performance is especially striking given that the photos were all taken on the same day, precluding changes in hairstyle, weight or health, and showed the face in frontal aspect under excellent lighting. The target photos were taken with a different camera to the array photos. This turns out to be an important factor in unfamiliar face recognition. The faces in the arrays were not chosen to be particularly homogeneous, and while they all met the same general description, this mirrors the real forensic situation (typically, police line-ups are populated with foils who broadly resemble the suspect).

In target-present arrays, participants failed to pick anyone on roughly 20 per cent of occasions, and on 10 per cent of occasions they picked the wrong person [22]. So, in the presence of the correct person, with a photo taken on the same day, in the same pose, in good light, people choose the wrong person from a line-up 10 per cent of the time. This is perhaps a surprising result. Participants were also willing to identify the wrong man in target-absent trials, doing so on 30 per cent of occasions. These results have now been replicated many times, and with different stimulus sets and observers [20].

More recent experiments have studied viewers' ability to make simple match/mismatch decisions to pairs of faces. These studies have been conducted using the same faces as used by Bruce *et al.* [24], as well as with Egyptian faces [20], and with a new set of faces which vary in age, gender and ethnicity. This latter set formed the basis of the Glasgow Face Matching Test (GFMT) [25], a psychometric instrument for measuring an individual's face matching ability (figure 1).

Using all these different stimulus sets, the same basic findings emerged: observers were very bad at matching pairs of unfamiliar faces, typically getting between 10 and 25 per cent of pairs wrong, even when viewing conditions were optimized in a way that could never be expected outside the laboratory. Participants were under no time constraints, worked in good lighting conditions, and were viewing high-quality photos that were taken on the same day. Merely using different cameras to capture the two photos is enough to impair performance. Photos sampled from the real world present an even greater challenge than those taken in research settings, as real world photos encompass the full natural range of variability. We refer to such photos as *ambient images*, to emphasize that they are sampled from the real world, rather than being posed photographs taken specifically for research purposes.

Figure 2 illustrates this problem. The top row shows photos of two *different* people, taken by the same photographer in the same town on the same day. Yet, it is difficult to see any basis for concluding that they are different people. By contrast, the bottom row shows photos of the *same* person, taken by the same photographer in the same room, approximately 18 years apart. In this case, it is difficult to see any basis for concluding that they are the same person.

If we add to this the fact that people may deliberately be trying to disguise their identities, we can see that the poor performance observed in laboratory studies almost certainly underestimates the applied problem. One very well-documented problem for face matching is the so-called 'other-race effect', in which viewers find it easier to recognize faces from their own race than faces from other races. In very recent work between universities in Egypt and Scotland, we have demonstrated the same phenomenon

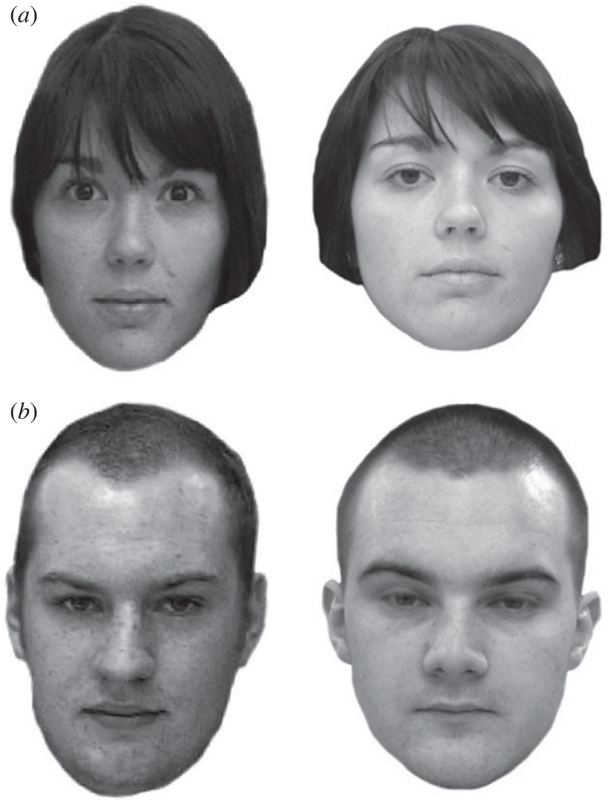

Figure 1. Two items from the GFMT— (*a*) a match, and (*b*) a mismatch. Performance on these items is surprisingly poor when the faces are unfamiliar.

in matching tasks: Egyptian viewers make more errors when matching Scottish faces, and the converse is true for Scottish viewers. Trying to match the face of someone from another race makes a bad situation worse: our experiments showed error rates of 20 per cent for one's own race, rising to about 25 per cent for matching faces from another race [26]. If a reader misidentified one in five words, we would not hesitate to say that they have difficulty in reading. Our ability to match unfamiliar faces is at that level. This presents a serious challenge for large-scale systems, where even a low percentage error rate can translate to thousands of individuals being misidentified. Consider that 200 000 people travel through Heathrow airport every day. In this setting, even 99 per cent accuracy would correspond to 2000 errors per day. There is nothing to suggest that anything approaching this level of accuracy is attainable in practice.

To get around these perceptual limitations, some practitioners working in the criminal justice system have sought objective measures of facial structure that could be used to match faces more reliably [27,28]. The basic approach, known as anthropometry, is to derive a numerical signature for each face by measuring the distances and angles between a small set of landmarks (e.g. the corners of the eyes, the centre of the mouth). Comparison of these standard metrics across images is then used to decide whether or not the images depict the same face. As it turns out, this approach is even less reliable than the normal visual inspection approach described above [29]. The reason anthropometry fails is that the small metric differences *between* faces are easily swamped by *within*-person changes in pose, expression and even the focal length of the camera lens. Simple geometric measures do not survive such changes, so images of different faces can easily give rise to more similar signatures than images of the same face.

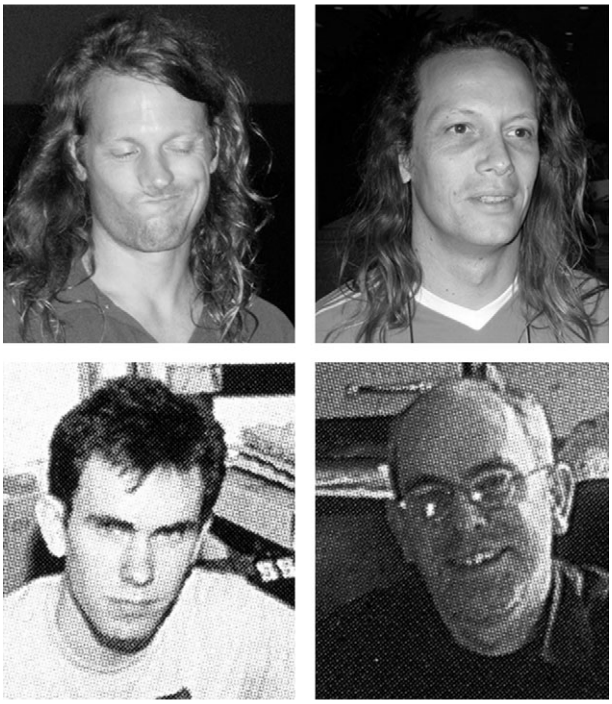

Figure 2. Real world photos of different people (top row) and the same person (bottom row). Bottom row photos from *20 years of Dischord* (2002), Washington, DC: Dischord Records. Reproduced with permission from Glen E. Friedman.

4. MISPLACED CONFIDENCE IN PHOTO-ID

The face matching results reviewed above show a level of performance that could certainly not be regarded as demonstrating expertise for faces. In this context, it is interesting to ask why our poor levels of performance seem to be so little understood. In practical settings, security experts and legislators continue to expand the use of photo-ID, asking inspectors to perform a task that is known to be highly error-prone. One possible reason is that photo-ID sounds intuitively plausible: we simply believe that we must be good at identifying people from photos. In this section, we consider why our intuition on this matter should be so at odds with the facts. We suggest that the basic problem is one of overgeneralization. Owing to the statistics of everyday social interaction, the idea of recognizing faces brings to mind particular face-processing tasks at which we really do excel. The problem arises when we assume that the same proficiency generalizes across all face-processing tasks.

One source of misplaced confidence in facial identification is that *image* recognition is often mistaken for *face* recognition, when the images happen to depict faces. Figure 3 illustrates this distinction. The question is the same for figure 1 above: do the two photos show the same person or different people?

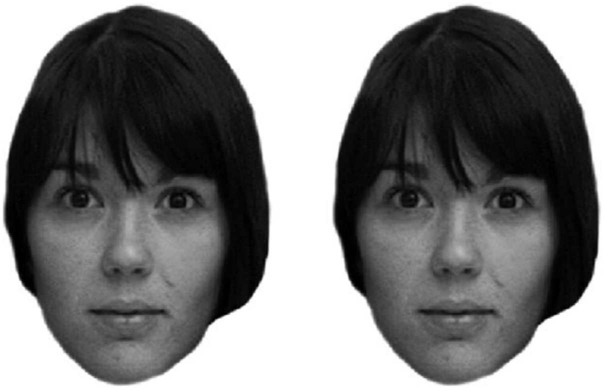

Figure 3. Face matching is easy with identical images, but that is not the applied problem.

This kind of pattern matching task can be solved with a line of computer code, and is completely trivial for human observers. Even honeybees (*Apis mellifera*) can solve a human recognition problem when the same image is used on successive presentations [30]. Image repetitions for faces are extremely common in daily life, so it is not surprising that our aptitude for spotting them is salient to us. Crucially, however, our facility with same image cases is irrelevant to applied face matching. This is because changes in the face itself, and the conditions of image capture, guarantee that *no face will give rise to the same image on any two occasions*. Confounding image matching and face matching therefore inspires false confidence. Image matching is easy where as *face* matching is difficult.

Perhaps the most compelling reason why people overestimate face-identification ability is that *familiar* face recognition is extremely good. Figure 4 shows a match task which is trivially easy for viewers who are familiar with Barack Obama's face.

The contrast with figure 1 is key. Even though unfamiliar face recognition can often be defeated by superficial image changes (e.g. a change in camera), familiar face recognition survives all manner of manipulations. Indeed identification of familiar faces remains highly accurate and robust, even when the quality of the image is severely degraded [31–33]. Burton *et al.* [31] found that students could match two images of their own lecturers almost perfectly, even when one of the images was a still taken from very poor-quality CCTV footage (figure 5).

Using exactly the same image pairs, viewers who were unfamiliar with the lecturers performed at chance levels. Familiarity does not merely improve performance—it completely transforms the task. This stark contrast in performance accords with neuropsychological and behavioural evidence for qualitative differences between unfamiliar and familiar face perception, including evidence from skin conductance studies [34–37], neuropsychological double dissociations [38–40], visual short-term memory capacity [41], analysis of information use [42] and individual differences [24,43].

We propose that non-psychologists addressing security issues are drawn to the use of face recognition because of our impressive ability to recognize *familiar* people. The mistake is to overgeneralize this expertise to *unfamiliar* faces. It is perhaps understandable that this overgeneralization should be so common. A great deal of the time that we spend looking at faces is spent looking at familiar faces, including those of family members, friends, colleagues and acquaintances, as well as media celebrities. Rather little of our contact time with faces involves people whom we have never seen before. Yet, virtually all of the applied interest in face recognition concerns that anomalous case. When an identity check is performed, it is generally to establish the identity of an unfamiliar person, rather than a familiar person. To complicate matters, insight into our overgeneralization is probably difficult to achieve, for at least two reasons. First, outside testing laboratories, we seldom receive feedback on our errors. If we encounter an unfamiliar person on one day, and then fail to recognize the same person the next day, we can simply assume that the second sighting was of a different person. In the absence of any feedback, this is a reasonable interpretation, but it leaves unchallenged the pernicious conviction that we never forget a face. Second, familiarity is rapidly acquired, so that we quickly leave behind our poor performance with new faces after rather modest exposure [44,45]. This brings us to the topic of face learning.

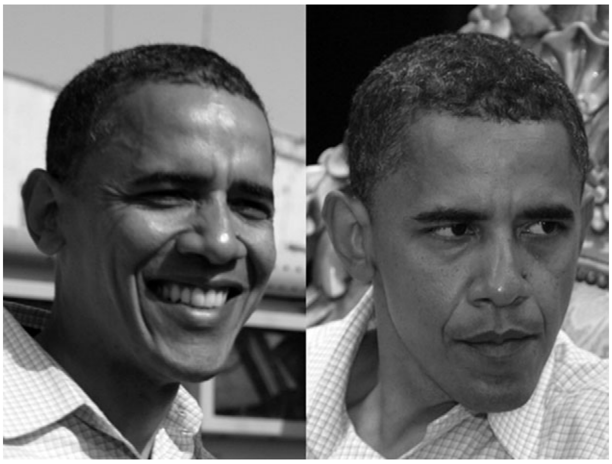

Figure 4. Two different photos of the same familiar person, taken on the same day. Familiar faces can be recognized across a wide range of images.

5. FACE LEARNING

We have made the case that there are important differences between familiar and unfamiliar face processing, which raises the question of how faces become familiar in the first place.

Every face that is familiar now was unfamiliar when it was first encountered, and has undergone a shift from being poorly recognized then to being well recognized now. Given the theoretical and applied significance of face learning, it is perhaps surprising how little research has been published on the topic. The most common approach to explaining the change in performance has been to posit a gradual development 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 face is learned. So, for

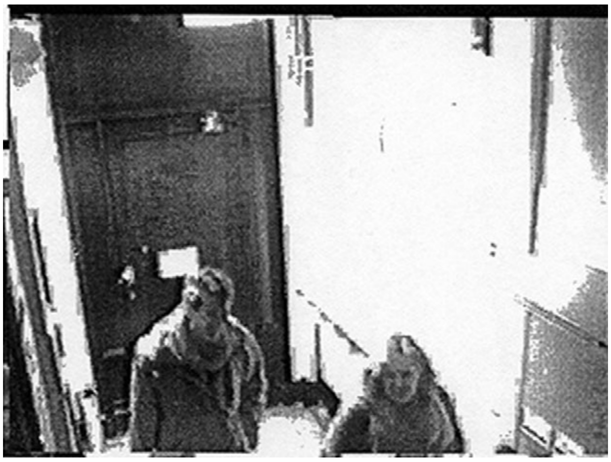

Figure 5. A CCTV image taken from an operational security camera at the University of Glasgow. Even these very poor-quality images are reliably recognized by viewers who are familiar with the faces.

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 [44–48]. Our own focus has been a very different proposal that does not involve an explicit shift in strategy, but instead emphasizes exposure-driven refinement of the stored representations against which seen 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. To examine how exposure might improve recognition, we developed the notion put forward by Bruce [49] of ‘stability from variation’. The thrust of this idea is that the variable nature of a person’s face allows the perceiver to distill a robust representation that incorporates aspects of appearance that are relevant for identification, while discarding the non-diagnostic information that is inherent in any particular set of images. In §6, we outline some limitations of photographic images for identification tasks before going on to describe our efforts to develop a more suitable image format.

6. LIMITATIONS OF PHOTOGRAPHIC IMAGES

The finding that human observers cannot reliably match unfamiliar faces raises an interesting issue. It could be that photographic face recognition is a *resource-limited* problem [50]. That is to say, the problem is solvable in principle—we just have not solved it yet. By this account, eventual success is just a matter of developing better procedures in the case of human performance, or better algorithms in the case of machine performance. This is presumably the conviction that has spurred the field on for some three decades. Alternatively, it could be that photographic face recognition is a *data-limited* problem [50]. That is to say, no amount of investment in matching procedures or matching algorithms will lead to useful levels of performance. This is a genuine possibility if performance is limited not by processing power or ingenuity, but by the information that is available in the image. There are plenty of problems outside face recognition that are data limited in this way (consider figure 6). Suppose we know that the central rectangle is the side elevation of a three-dimensional solid. If that is all the information we have, there is no computation that we can perform which will allow us to establish whether the solid is a cylinder or a rectangular block. There simply is not enough information in the image to allow us to distinguish among those possibilities. To take an example from another domain, consider the word ‘bank’. We can analyse this arrangement of letters for as long as we like. The letters alone will never reveal whether the referent is a financial institution or the side of a river.

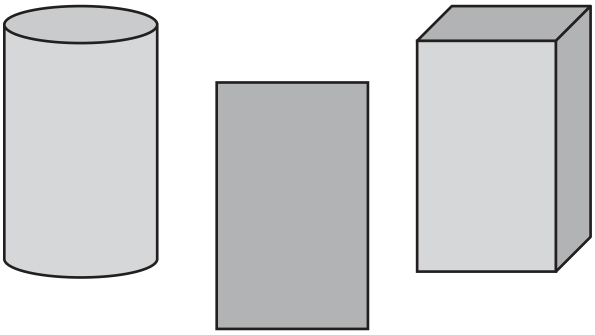

Figure 6. A data-limited problem. The central rectangle does not contain enough information to disambiguate the object. Identification of faces from two- or three-dimensional snapshots could also be a data-limited problem.

We propose that it is worth entertaining the possibility that unfamiliar face matching is a data-limited problem. It might not be possible *in principle* to achieve useful levels of accuracy when matching pairs of photographs if face photographs do not contain enough information to disambiguate identity.

If matching unfamiliar faces is a data-limited problem, one response would be to stop trying to match *photographs*, and instead try to develop alternative face representations that are better suited to the task. Given the recency of portrait photography in human evolution, there is little reason to expect that the human visual system should be well-suited to processing facial ‘snapshots’, as these only occur in the context of photography. In natural samples, the facial image changes from moment to moment, as well as from year to year. This variability provides an opportunity to separate aspects of appearance that are common across all the images (and hence potentially diagnostic of identity), from those that are transient (and hence specific to a particular image). As a single photograph fuses these independent streams of information, it does not allow the viewer to separate the *face* from the *image*. Note that increasing the resolution of the image, or capturing it in three-dimensions, are not helpful in this respect. We propose that if we are to match the face reliably, it will be necessary to develop stable representations that are not dominated by image transients.

7. STABLE FACE REPRESENTATIONS

Our own research on stable face representations has focused on a very simple proposal based on averaging together several photos of the same face. This concept

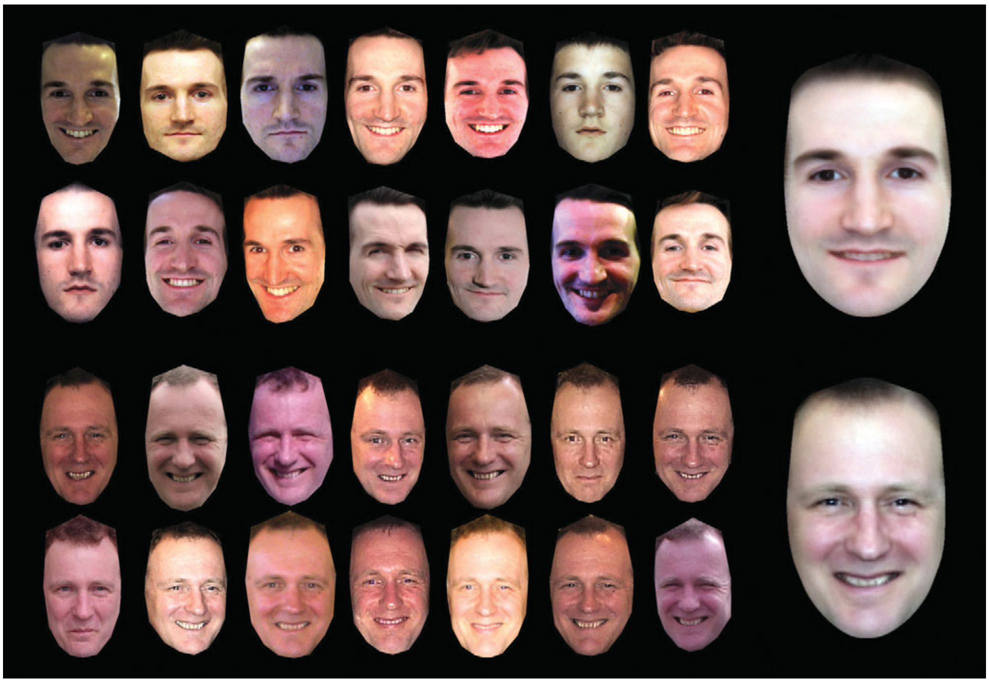

Figure 7. Average images (right) and their constituent photographs (left) for authors R.J. (top row) and A.M.B. (bottom row).

was introduced by Galton [51], and developed by Benson & Perrett [52]. We have extended the original method in order to investigate theoretical and practical aspects of face recognition [53–56]. In this technique, multiple photographs of each face are collected from existing sources (e.g. the Internet). These ambient images are intended to capture a natural range of variation in facial appearance.

To construct the average image, the facial shape for each image is first captured by recording the *xy*-coordinates of multiple facial landmarks (e.g. corners of the eyes, tip of the nose). This step is performed manually by a human operator, and also has the effect of segmenting the face region from the background. The landmarked images are then co-registered by morphing them to a standard template using bi-cubic interpolation. For each face, we derive the average texture from the co-registered images by calculating the mean intensity values at each pixel, and the average shape of the corresponding unregistered images by calculating the mean *xy*-coordinates of each facial landmark. We then morph each person's average texture to their average shape to produce the stabilized image of their face. Because non-diagnostic information such as lighting direction is uncorrelated with identity, for each person it regresses to the mean. The process thus dilutes aspects of the image that change from one photo to the next, while preserving aspects of the image that are consistent across the set. Figure 7 shows the results of this process applied to photos of the authors.

We have previously presented a simple simulation that explains why this process should improve identification accuracy [53]. In the simulation, each face image is represented as a point in multi-dimensional space. Some of the dimensions code physical differences between the faces of individuals. Others code differences in lighting, pose and other image-level factors. As we have seen, the problem that plagues recognition of faces is that image variability tends to outweigh face variability, at least at the pixel level, so that photos of different faces can be more similar than photos of the same face. We modelled these changes with two sets of Gaussian random variables. The first set models physical differences between individuals. The variance of these variables is relatively small, as all faces share the same basic anatomy, but the average value will be non-zero for any given person, as individual faces are different nonetheless. The second set of variables models variations due to image artefacts such as lighting and pose. These have a relatively high variance, owing to their wide natural range, but a mean of zero, reflecting the fact that for any face, lighting is equally likely to come from any direction. Our main interest was the application of principal components analysis (PCA) to this model. As described, most of the variance lies along the image dimensions, so that is what PCA picks up. The early components thus code mainly image-level variability, which is irrelevant to the task of identification. However, if the different examples of each face are averaged together prior to PCA, much of the image variance will be eliminated. The analysis will then be forced to code face dimensions, which are the ones that are interesting for recognition purposes.

Figure 8 shows the results of the simulation. As the standard deviation of the image components increases, the hit rate for new exemplars falls rapidly. However, when the images are averaged before the PCA, performance is much more resilient. We now consider the implications of this procedure in the context of real face images.

8. PROPERTIES OF AVERAGE IMAGES

When applied to actual face photographs, the same logic gives rise to some interesting properties that are

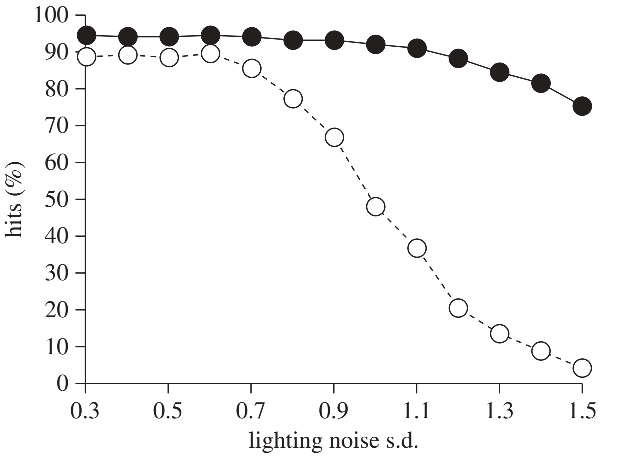

Figure 8. Average hit rates for average (filled circles) and exemplar (open circles) models, under varying amounts of image noise. The model based on average images exhibits greater tolerance to noise.

highly desirable from an identification standpoint. First, the average stabilizes surprisingly quickly [53,54]. Even when only two photographs of the face are available, identification performance is better when these are averaged together than when they are treated separately. The average becomes increasingly stable as subsequent photographs refine it further. By the time about a dozen photographs have been incorporated, the image has more or less settled, and adding further photographs makes little difference. A stable average thus emerges quickly. Secondly, it does not matter which particular photographs of the person are used—averages based on any set of ambient images of the person look much the same. In other words, the average constructed from photos one to 10 of a given face converges on the average constructed from photos 11 to 20. This is a very useful property when it comes to sharing results from different systems, as it means the systems are not required to work from the same source images. We have investigated these properties across a number of studies [54].

For each of six different faces in our initial study, we computed a reference image by averaging together 32 ambient photographs. We then computed the following additional averages, based on subsets of each individual's face: sixteen 2-image averages, eight 4-image averages, four 8-image averages and two 16-image averages. Within any n-image level, each constituent photograph contributed to only one average image. For each face, we then computed the pixel-wise difference between each subset average and the reference image. An example of this analysis is shown in figure 9. For all identities, the four 8-image averages were already very similar, despite the fact that each average was constructed from completely independent sets of ambient photos.

A third attractive property of the averaging process is that it incorporates robustness against errors. Errors must be considered inevitable in a large database, so it is essential that the system does not collapse when errors arise. Image averaging can deliver representations that are highly resistant to contamination from misidentified photographs. As long as the average is constructed from enough images, recognition is barely affected when a few photos come from a different person, as illustrated in figure 10 where column (*a*) shows 'pure' average images for two different individuals (Leonardo DiCaprio and Brad Pitt). In each case, the average is constructed from 20 photographs. Column (*b*) shows average images of the same two individuals, this time contaminated with misclassified photographs.

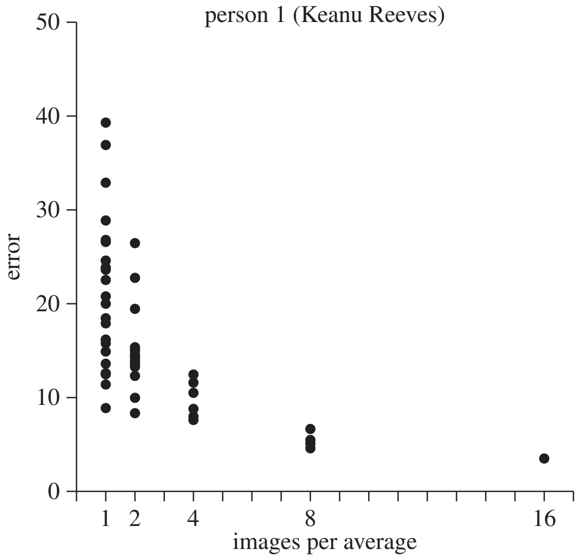

Figure 9. Pixel-wise error between reference image, and averages constructed from subsets. Error reduces sharply as more photographs contribute to the average, indicating rapid convergence.

Two different types of error are shown. The top row illustrates systematic misidentification, in which the average is composed of 16 photos of the base individual (Leonardo DiCaprio) and four photos of an intruder. The bottom row illustrates a similar example involving random misidentification. Here the four intruding images are from four different males. Even at this high contamination rate (20%), the identity signal is robust. The contaminated averages on the right are easy to identify as the individuals in the left column. Indeed it is difficult to see that anything is amiss with these images. At the same time, the identities of the intruding individuals are virtually impossible to discern. Consistent with these observations, we have found in simulations that face recognition performance undergoes graceful degradation rather than catastrophic collapse as the level of contamination in the average images is increased [54].

9. HUMAN FACE RECOGNITION BASED ON AVERAGE IMAGES

As we have seen, the simple process of averaging together photographs of a face forms a stable representation of its appearance. We propose that such representations may be capable of supporting generalized face identification that is less image bound, as found in observers who are familiar with the face. Moreover, the refinement of the average, driven by increased exposure, may provide a useful model of face learning. Evidence for this comes from behavioural

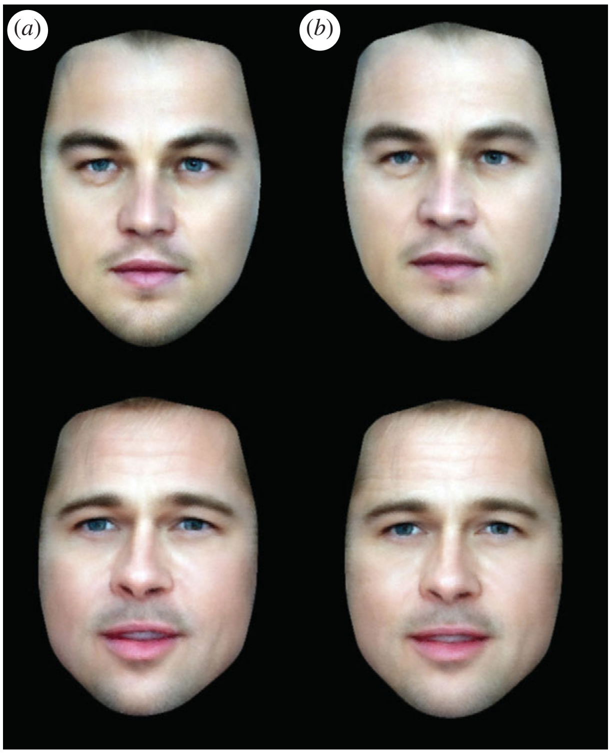

Figure 10. (*a*) Average images are robust against contamination from photographs of the wrong person (*b*). The top right image was constructed from 16 photos of Leonardo DiCaprio and four photos of George W. Bush. The bottom right image was constructed from 16 photos of Brad Pitt, and one photo for each of Bill Clinton, Jack Nicholson, John Travolta and Tom Cruise.

experiments in which we compare recognition performance for famous faces presented in average image or standard photograph format [53]. In name verification tasks, observers are presented with a famous name followed by a face. Their task is to press one key if the face matches the preceding name, and another key if the face belongs to someone else. In our experiments, the face was equally likely to be presented as a standard photograph or as an average image. We assumed that face recognition involves matching the seen face to a representation stored in memory, and that the speed of recognition indexes the closeness of that match, with faster reaction times indicating closer correspondence. Figure 11 shows the results from one such study. Whether correctly accepting a match, or correctly rejecting a mismatch, responses were faster for average images than for photographs.

In a separate study, we found that identification of average images becomes increasingly efficient as more and more photos contribute to the average. The graded benefit seen here for incorporating increased exposure into the representation echoes the graded benefit in face matching as the faces are learned [57–59]. Taken together, these behavioural findings provide quite compelling support for the notion that an average image of an individual's face is a relatively good match to a familiar observer's mental representation of that face, compared with a photograph. Given that perception of familiar faces is already extremely efficient, even when based on photographs, it may be surprising that any representation can improve on this performance. The average nonetheless seems

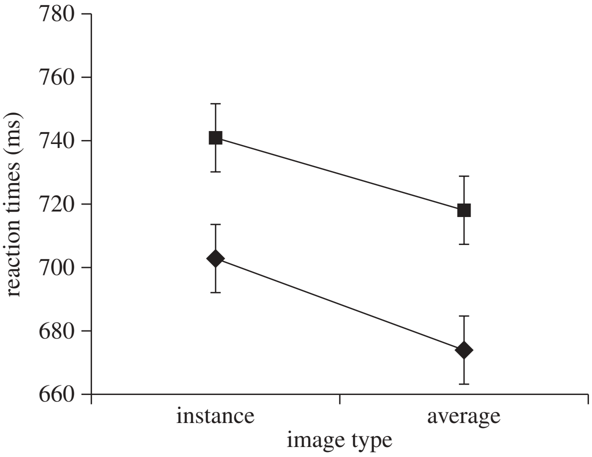

Figure 11. Mean reaction times in a speeded name verification task. Average images were processed more efficiently than photographs in this task. Filled squares, invalid cue; filled diamonds, valid cue.

to capture the essence of a person's appearance in a way that facilitates identification.

10. AUTOMATIC FACE RECOGNITION BASED ON AVERAGE IMAGES

In earlier sections, we argued that automatic face recognition systems can be unreliable because they model a process that is unreliable in humans—specifically, matching photos of unfamiliar faces. Given that familiarity yields robust face recognition in human observers, we have investigated whether image averaging, as a model of familiarity, can similarly improve the performance of machine systems.

The general approach to automatic face recognition is to compare a previously unseen 'probe' image against a gallery of 'enrolled' images stored in memory. If a sufficiently close match is found, then a hit is recorded. To date, research into automatic face recognition systems has focused almost exclusively on improving the matching algorithm [17,18]. The fact that probe and gallery images must be matched across a range of viewing conditions poses a difficult problem. It requires a good understanding of the illumination, the reflectance properties of skin and hair, the location and optics of the camera, and an assumption that the physical appearance of the person is relatively unchanged between capture of the gallery and probe images. The conventional approach to this problem is to try to 'partial out' aspects of the image that are not specific to the individual, before attempting a match. Our approach has been very different. Instead of focusing on the matching algorithm, we have considered the representation of the face. It is in this context that the benefit of image averaging becomes clear. Across various matching algorithms, the match between a photo of a face and an average image of a the same face is generally closer than the match between two photos of the same face. Moreover, the performance of averages improves as more photos are incorporated into it. Figure 12 illustrates this with data from a PCA study [53].

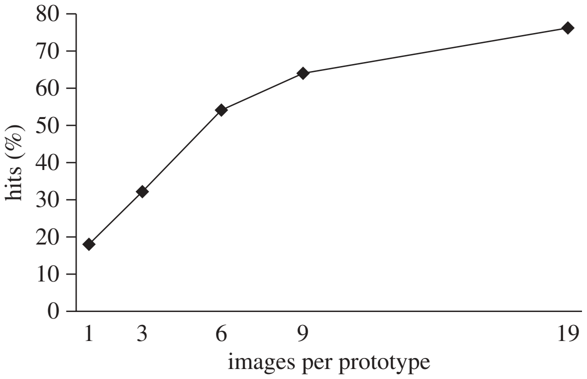

Figure 12. Face recognition accuracy in a PCA-based model. Accuracy increases with the number of photos contributing to each average image.

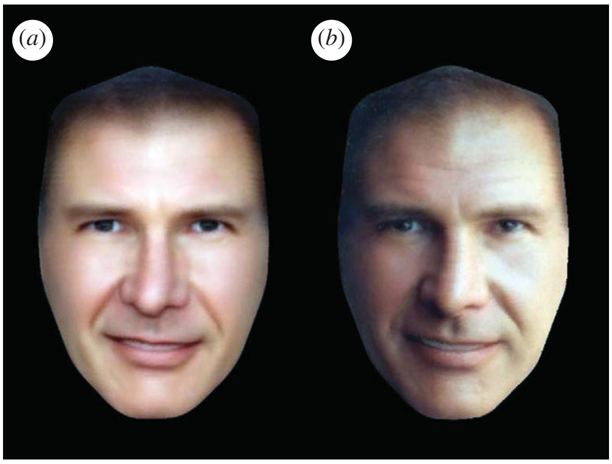

Figure 13. (*a*) An average image of Harrison Ford, and (*b*) the same average image with realistic illumination reintroduced.

In this study, identity averages were generated from 3, 6, 9 or 19 ambient images of each person, with a 20th image (selected at random) serving as the probe in all these conditions. Single photographs of the faces were also enrolled for comparison. The resulting five image formats were counterbalanced with respect to identity, so that across the whole study, each face was represented by each type of image. The nearest neighbour match for each probe was then computed using a Mahalanobis distance metric [60]. As can be seen from figure 12, the hit rate improves as more images of each face are averaged together. This rather dramatic improvement seems to capture well the advantage of familiarity: the more encounters an observer has had with a person (modelled here as the more images which are incorporated into the average), the more reliable subsequent identification becomes. This pattern of graded improvement in the model mirrors the pattern seen in the behavioural studies described above. The scale of the gain is also worth noting. Overall hit rate increased from 18 per cent when matching pairs of photos, to 75 per cent when matching photos to 19-image averages. This is a considerable improvement. Note that the matching algorithm and the probe images were identical in these conditions, and that the only component of the system that changed was the format of the gallery images.

More recently [55,56], we have replicated the same basic pattern using a completely different matching algorithm based on wavelet decomposition [61]. For these studies, we used an online implementation of FaceVACS, an industry standard face recognition system that has been used in airport security deployments (e.g. SmartGate). Importantly, we had no control over the calibration of the matching algorithm, or the associated gallery of over 30 000 enrolled images. Nonetheless, the system performed much better with average images than with photographs. Indeed, the averaging process raised the hit rate from 54 to 100 per cent. Even when the averages were constructed entirely from unrecognizable photographs (i.e. 0% hits), the hit rate for the average images was 80 per cent. Such performance levels are unprecedented for images with realistic variability, and suggest that image averaging may be worth pursuing as a means of attaining the robust performance of a familiar observer in machine systems. As the image gallery that we used in that study is constantly expanding, we were able to estimate test–retest reliability by resubmitting the same probes to an even larger gallery a year later [56]. Once again, photos performed poorly, but the average images were all correctly identified. Interestingly, some of the averages were now matched to newly enrolled photos of the correct person, underscoring the generality of the average-to-photo mapping.

One pragmatic advantage of this approach is that it requires only a change in the images that are used, and not a change in the processing that the images undergo. This is an important feature in the context of automatic face recognition: our proposal does not compete with previous advances in automatic face recognition research. On the contrary, it makes an independent contribution that can complement existing algorithms for additive benefit [55].

11. SUMMARY AND FUTURE DIRECTIONS

In this review, we have emphasized the importance of representations when it comes to matching faces. We conclude that photography is not necessarily a good way to capture facial appearance. In contrast, we have been surprised at the apparent promise of a very simple representation-based summary statistics. Across a range of studies using familiar faces, we have shown that observers find these average images easier to identify than the constituent photographs, as indexed by faster reaction times in a name verification task, and higher hit rates in a spontaneous naming task [53]. We interpret these findings as evidence that the average image is a relatively close match to the observer's mental representation of that face, compared with a photo. Automatic face recognition systems can also benefit greatly from using average images instead of photographs, even when the matching algorithm is held constant [53–56]. As in the human case, eliminating image artefacts means that they cannot dominate the match. Image averaging is by no means the only way to accomplish this goal, and we certainly do not claim that it is the best. It is simply one attempt to integrate a psychological model of face learning into the automatic case. Many issues remain outstanding. We therefore finish by noting some practical constraints, and highlighting some promising avenues for further development of this approach.

The first point concerns not the representation itself, but the steps involved in its construction. At present, the averaging process is limited by a bottleneck in landmarking the constituent photographs. This step is time-consuming, as it is carried out manually by an operator who locates individual landmarks in each photograph sequentially. It would clearly be advantageous to be able to automate this step. Although there has been considerable progress in automatic landmarking in recent years, reliable landmarking is very difficult when using ambient images. As such, it cannot yet be accomplished automatically. This is certainly a practical constraint, but note that it is independent of our theoretical claims: it is easy to locate landmarks on a face that you cannot recognize, and doing so does not trigger identification. This dissociation demonstrates that face landmarking and face identification rely on distinct processes. Nevertheless, automatic landmarking would be extremely useful. It would allow the construction of average images to run unsupervised, thereby speeding the whole process, and allowing stable representations to be derived by automated systems. Without such an advance, it is not clear how image averaging could be deployed for use in border control or other large-scale operations. It is possible that a running average of the passport holder's face could be updated with a new photograph every time the passport is used. Alternatively, the technique may be better suited to more targeted use, as when monitoring for particular individuals.

A second observation concerns the somewhat uncanny appearance of average images. The power of image averaging is that it washes out illumination and other artefacts that could otherwise mislead the match. An incidental consequence of this is that the images have visual characteristics could not occur in the real world. Their appearance has been described as 'unreal' and 'too good to be true' by experimental participants. Such impressions may be revealing. The more closely an image comes to approximate the statistical average of an individual's face, the less closely it approximates an observer's experience of that face, which necessarily incorporates the kinds of environmental noise that image averaging removes. It is possible that systematically reintroducing some of this noise could make the average image look more real, without compromising its stability (figure 13). This is a possibility that we are beginning to explore [62].

A third issue concerns limitations of the average as a summary statistic. As we have described, our computation of the average face is based on the arithmetic means of pixel intensity values and *xy*-coordinates of facial landmarks. The rationale for calculating the average is simply that it summarizes a variable set of values by estimating the central tendency. We have shown that this operation alone can be useful in boosting identification performance. However, in virtually any statistical situation, central tendency is rather uninformative by itself. It is much more useful when combined with information about the variability of the set. Very recently, we have begun systematically to analyse variability in photos of the same face. Our early results suggest that this will prove to be a very fruitful line of inquiry [62]. For now, we conclude that the appearance of an individual face is inherently variable, and that robust identification requires this variability to be stabilized.

This research was supported by ESRC grants to R.J. and A.M.B. (RES-062-23-0549 and RES-000-22-2519).

REFERENCES

1 Clifford, B. R. & Bull, R. 1978 *The psychology of person identification.* London, UK: Routledge & Kegan Paul.

2 Loftus, E. F. 1979 *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

3 Yarmey, A. D. 1979 *The psychology of eye-witness testimony.* New York, NY: The Free Press.

4 Wells, G. L. & Olson, E. A. 2003 Eyewitness testimony. *Annu. Rev. Psychol.* **54**, 277–295. (doi:10.1146/annurev.psych.54.101601.145028)

5 Sporer, S. L. 2008 Lessons from the origins of eyewitness testimony research in Europe. *Appl. Cogn. Psychol.* **22**, 737–757. (doi:10.1002/acp.1479)

6 Morton, J. 1969 Interaction of information in word recognition. *Psychol. Rev.* **76**, 165–178. (doi:10.1037/h0027366)

7 Morton, J. 1979 Facilitation in word recognition: experiments causing change in the language model. In *Processing visible language*, vol. 1 (eds P. A. Kolers, M. Wrolstad & H. Bouma), pp. 259–268. New York, NY: Plenum.

8 Seymour, P. H. K. 1979 *Human visual cognition.* London, UK: Collier Macmillan.

9 Nelson, D. L., Reed, V. S. & MacEvoy, C. L. 1977 Learning to order pictures and words: a model of sensory and semantic encoding. *J. Exp. Psychol. Hum. Learn. Mem.* **2**, 49–57. (doi:10.1037/0278-7393.2.1.49)

10 Warren, C. & Morton, J. 1982 The effects of priming on picture recognition. *Br. J. Psychol.* **73**, 117–129.

11 Bruce, V. & Young, A. W. 1986 Understanding face recognition. *Br. J. Psychol.* **77**, 305–327.

12 Ellis, H. 1975 Recognizing faces. *Br. J. Psychol.* **66**, 409–426.

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

14 Kanade, T. 1977 *Computer recognition of human faces.* Basel and Stuttgart: Birkhauser.

15 Stonham, I. J. 1986 *Practical face recognition and verification with WISARD: aspects of face processing.* Dordrecht, The Netherlands: Martinus Nijhoff.

16 Kohonen, T., Oja, E. & Lehtio, P. 1981 *Storage and processing of information in distributed associative memory systems, in parallel models of associative memory* (ed. J. A. Anderson), pp. 105–143. Hillsdale, NJ: Lawrence Erlbaum Associates.

17 Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L. & Sharpe, M. 2007 FRVT 2006 and ICE 2006 large-scale results. Technical Report NISTIR 7408. National Institute of Standards and Technology, Gaithersburg, MD, USA.

18 Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L. & Sharpe, M. 2010 FRVT

2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 831–846. (doi:10.1109/TPAMI.2009.59)

19 Kemp, R., Towell, N. & Pike, G. 1997 When seeing should not be believing: photographs, credit cards and fraud. *Appl. Cogn. Psychol.* **11**, 211–222. (doi:10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O)

20 Megreya, A. M. & Burton, A. M. 2008 Matching faces to photographs: poor performance in eyewitness memory (without the memory). *J. Exp. Psychol. Appl.* **14**, 364–372. (doi:10.1037/a0013464)

21 Davis, J. & Valentine, T. 2009 CCTV on trial: matching video images with the defendant in the dock. *Appl. Cogn. Psychol.* **23**, 482–505. (doi:10.1002/acp.1490)

22 Bruce, V., Henderson, Z., Greenwood, K., Hancock, P., Burton, A. M. & Miller, P. 1999 Verification of face identities from images captured on video. *J. Exp. Psychol. Appl.* **5**, 339–360. (doi:10.1037/1076-898X.5.4.339)

23 Bruce, V., Henderson, Z., Newman, C. & Burton, A. M. 2001 Matching identities of familiar and unfamiliar faces caught on CCTV images. *J. Exp. Psychol. Appl.* **7**, 207–218. (doi:10.1037/1076-898X.7.3.207)

24 Megreya, A. M. & Burton, A. M. 2006 Unfamiliar faces are not faces: evidence from a matching task. *Mem. Cogn.* **34**, 865–876. (doi:10.3758/BF03193433)

25 Burton, A. M., White, D. & McNeill, A. 2010 The Glasgow Face Matching Test. *Behav. Res. Methods* **42**, 286–291. (doi:10.3758/BRM.42.1.286)

26 Megreya, A. M., White, D. & Burton, A. M. Submitted. Perceptual aspects of the other race effect: evidence from a matching task.

27 Vanezis, P., Lu, D., Cockburn, J., Gonzalez, A., McCombe, G., Trujillo, O. & Vanezis, M. 1996 Morphological classification of facial features in adult Caucasian males based on an assessment of photographs of 50 subjects. *J. Forensic Sci.* **41**, 786–791.

28 Porter, G. & Doran, G. 2000 An anatomical and photographic technique for forensic facial identification. *Forensic Sci. Int.* **114**, 97–105. (doi:10.1016/S0379-0738(00)00290-5)

29 Kleinberg, K. F., Vanezis, P. & Burton, A. M. 2007 Failure of anthropometry as a facial identification technique using high quality photographs. *J. Forensic Sci.* **52**, 779–783. (doi:10.1111/j.1556-4029.2007.00458.x)

30 Dyer, A. G., Neumeyer, C. & Chittka, L. 2005 Honeybee (*Apis mellifera*) vision can discriminate between and recognize images of human faces. *J. Exp. Biol.* **208**, 4709–4714. (doi:10.1242/jeb.01929)

31 Burton, A. M., Wilson, S., Cowan, M. & Bruce, V. 1999 Face recognition in poor quality video: evidence from security surveillance. *Psychol. Sci.* **10**, 243–248. (doi:10.1111/1467-9280.00144)

32 Harmon, L. D. 1973 The recognition of faces. *Sci. Am.* **227**, 71–82.

33 Sergent, L. 1986 Microgenesis of face perception. In *Aspects of face processing* (eds H. D. Ellis, M. A. Jeeves, F. Newcombe & A. W. Young). Dordrecht, The Netherlands: Martinus Nijhoff.

34 Tranel, D. & Damasio, A. R. 1985 Knowledge without awareness: an autonomic index of facial recognition by prosopagnosics. *Science* **228**, 1453–1454. (doi:10.1126/science.4012303)

35 Bauer, R. M. 1984 Autonomic recognition of names and faces in prosopagnosics: a neuropsychological application of the Guilty Knowledge Test. *Neuropsychologia* **22**, 457–469. (doi:10.1016/0028-3932(84)90040-X)

36 Ellis, H. D., Young, A. W. & Koenken, G. 1993 Covert face recognition without prosopagnosia. *Behav. Neurol.* **6**, 27–32.

37 Ellis, H. D., Quayle, A. H. & Young, A. W. 1999 The emotional impact of faces (but not names): face specific changes in skin conductance responses to familiar and unfamiliar people. *Curr. Psychol.* **18**, 88–97. (doi:10.1007/s12144-999-1018-y)

38 Benton, A. L., Hamsher, K. S., Varney, N. R. & Spreen, O. 1983 *Contributions to neuropsychological assessment.* New York, NY: Oxford University Press.

39 Young, A. W., Newcombe, F., de Haan, E. H. F., Small, M. & Hay, D. C. 1993 Face perception after brain injury: selective impairments affecting identity and expression. *Brain* **116**, 941–959. (doi:10.1093/brain/116.4.941)

40 Malone, D. R., Morris, H. H., Kay, M. C. & Levin, H. S. 1982 Prosopagnosia: a double dissociation between the recognition of familiar and unfamiliar faces. *J. Neurol. Neurosurg. Psychiatry* **45**, 820–822. (doi:10.1136/jnnp.45.9.820)

41 Jackson, M. C. & Raymond, J. E. 2008 Familiarity enhances visual working memory for faces. *J. Exp. Psychol. Hum. Percept. Perform.* **34**, 556–568. (doi:10.1037/0096-1523.34.3.556)

42 Osborne, C. D. & Stevenage, S. V. 2008 Internal feature saliency as a marker of familiarity and configural processing. *Vis. Cogn.* **16**, 23–43. (doi:10.1080/13506280701238073)

43 Megreya, A. M. & Burton, A. M. 2007 Hits and false positives in face matching: a familiarity-based dissociation. *Percept. Psychophys.* **69**, 1175–1184. (doi:10.3758/BF03193954)

44 Bonner, L., Burton, A. M. & Bruce, V. 2003 Getting to know you: how we learn new faces. *Vis. Cogn.* **10**, 527–536. (doi:10.1080/13506280244000168)

45 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. (doi:10.1068/p080431)

46 Kaufmann, J. M., Schweinberger, S. R. & Burton, A. M. 2009 N250 ERP correlates of the acquisition of face representations across different images. *J. Cogn. Neurosci.* **21**, 625–641. (doi:10.1162/jocn.2009.21080)

47 O'Donnell, C. & Bruce, V. 2001 Familiarisation with faces selectively enhances sensitivity to changes made to the eyes. *Perception* **30**, 755–764. (doi:10.1068/p3027)

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. (doi:10.1068/p140737)

49 Bruce, V. 1994 Stability from variation: the case of face recognition. The M. D. Vernon memorial lecture. *Q. J. Exp. Psychol.* **47A**, 5–28.

50 Norman, D. A. & Bobrow, D. G. 1975 On data-limited and resource-limited processes. *Cogn. Psychol.* **7**, 44–64. (doi:10.1016/0010-0285(75)90004-3)

51 Galton, F. 1878 Composite portraits. *J. Anthropol. Inst.* **8**, 132–144.

52 Benson, P. J. & Perrett, D. I. 1993 Extracting prototypical facial images from exemplars. *Perception* **22**, 257–262. (doi:10.1068/p220257)

53 Burton, A. M., Jenkins, R., Hancock, P. J. B. & White, D. 2005 Robust representations for face recognition: the power of averages. *Cogn. Psychol.* **51**, 256–284. (doi:10.1016/j.cogpsych.2005.06.003)

54 Jenkins, R., Burton, A. M. & White, D. 2006 Face recognition from unconstrained images: progress with prototypes. *In Proc. 7th IEEE Int. Conf. on Automatic Face and Gesture Recognition, Southampton, UK, 2–6 April 2006*, pp. 25–30. (doi:10.1109/FGR.2006.45)

55 Jenkins, R. & Burton, A. M. 2008 100% accuracy in automatic face recognition. *Science* **319**, 435. (doi:10.1126/science.1149656)

56 Jenkins, R. & Burton, A. M. 2008 Response to comment on '100% accuracy in automatic face recognition'. *Science* **321**, 912. (doi:10.1126/science.1158428)

57 Clutterbuck, R. & Johnston, R. A. 2002 Exploring levels of face familiarity by using an indirect face-matching measure. *Perception* **31**, 985–994. (doi:10.1068/p3335)

58 Clutterbuck, R. & Johnston, R. A. 2004 Matching as an index of face familiarity. *Vis. Cogn.* **11**, 857–869. (doi:10.1080/13506280444000021)

59 Clutterbuck, R. & Johnston, R. A. 2005 Demonstrating how unfamiliar faces become familiar using a face matching task. *Eur. J. Cogn. Psychol.* **17**, 97–116. (doi:10.1080/09541440340000439)

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

61 Wiskott, L., Fellous, J.-M., Kruger, N. & von der Malsburg, C. 1997 Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **17**, 775–779. (doi:10.1109/34.598235)

62 Burton, A. M., Jenkins, R. & Schweinberger, S. R. In press. Mental representations of familiar faces. *Br. J. Psychol.*