

Limitations in Facial Identification: The Evidence

Dr R. JENKINS*, Professor A. M. BURTON*

The police shooting of Jean Charles de Menezes in 2005, underscored the importance of accurate face identification, and the tragic consequences of getting it wrong. That was a particularly high profile case, but there are many cases of mistaken identity every year. Some of these may lead to no more than inconvenience or frustration, but others can lead to innocent suspects being convicted or to guilty suspects being freed. The scale of this problem becomes evident when DNA testing reveals prisoners who were convicted on the basis of face identification to be innocent (see, for example, <http://www.innocenceproject.org/understand/Eyewitness-Misidentification.php>).

It has long been understood that juries find eyewitness identification compelling, despite its fallibility (Wells, 1993; Wells *et al*, 2000). One possible reason for our faith in eyewitness identification is that we are so good at recognizing familiar faces in everyday life that we tend to over-generalize our confidence in these abilities, by assuming that we are also good at recognizing unfamiliar faces. This is a mistake. Research into the psychology of face perception has led to great progress in understanding the complexity of everyday face recognition. In doing so, it has also revealed fundamental limitations in the face recognition abilities of both humans and machines. These limitations have profound implications for today's surveillance society (Wood, 2006). With more than 4m CCTV cameras operating in the UK, photo identity cards in the pipeline, and national security at the top of the political agenda, it has never been more important to understand these limitations. In this article, we consider the strengths and weaknesses of human face recognition with reference to experimental evidence. We also show how machine performance can be improved by incorporating discoveries from psychological research.

National security, crime prevention, and access to services and resources, often depend on our ability to establish the identity of an individual and check that they are who they claim to be. This proof of identity is most frequently achieved by comparing the individual's appearance to a previously captured image. This may be a passport photograph in the case of immigration, or a still from CCTV footage in forensic settings. However, research has shown that it is surprisingly difficult to match a face to an image. This routine task, performed hundreds of times every day by passport officers, security guards and police officers, turns out to be highly error prone. In one of the first demonstrations of this phenomenon Kemp, Towell & Pike (1997) undertook 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 required 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 they were taking part in a study concerning the utility of photo-credit cards, they performed surprisingly badly, with about half of the fraudulent cards being accepted, and about one in 10 of the valid cards being falsely rejected. The situation is no better when viewers are asked to compare photographs. In a striking demonstration of this, Bruce *et al* (1999, 2001) 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 alongside a face target. All the images were of clean-shaven young men, and all were taken on the same day, but the target and array photos were taken with different high-quality cameras. Again, viewers were strikingly poor at this task. Over a series of experiments, they were incorrect a quarter of the time, despite having no time-pressure, no memory load, and same-day photos. Even when the task was reduced to a simple two-item match (do these two photos show the same person or not?) the error rate was one in five, (Megreya & Burton, 2006, 2007), a surprisingly low level of performance. In recent follow-up work, we were able to compare people's ability to match two photos, with their ability to match a photo to a real face. Using the same target individuals (all young men), we again found error rates of one in five or worse. Viewers were no better at matching the real face to a photo than they were at matching two photos.

The poor levels of performance from these experiments actually show the best results that could be achieved by viewers. In these experiments the participants were under no time constraints, worked in good lighting conditions, and were viewing high quality photos taken on the same day. In realistic settings, people's appearance changes, even day to day. If we add to this the fact that people may deliberately be trying to disguise their identities, may change their hair, put on weight, or suffer ill-health, we can see that the poor performance observed in our experiments almost certainly underestimates the real world problem. One very common problem for face matching is the well-documented "other-race effect", in which viewers find it easier to recognize faces from their own race than faces from other races. In some recent work between universities in Egypt and Scotland, we have observed the same phenomenon for matching photos: 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 one in five for one's own race, rising to about one in four for matching faces from another race. If a reader misidentified one in five words, we would not hesitate to say that they have difficulty reading. Our ability to match unfamiliar faces is at that level. This presents a particularly serious problem

* Department of Psychology, University of Glasgow, Glasgow, UK.
E-mail: rob@psy.gla.ac.uk, Tel: +44 (0) 141 330 4663, Fax: +44 (0) 141 330 4606

in large scale systems, where even a low percentage error rate can translate to thousands of misidentifications. Consider that 200,000 people travel through Heathrow airport every day. In this setting, even 99 *per cent* accuracy would correspond to 2,000 errors *per day*.

Confronted with these limitations in face matching performance, some practitioners have sought objective measures of facial structure that could individuate faces in a more reliable fashion (eg, Vanezis *et al*, 1996; Porter & Doran, 2000). 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 (eg, 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 visual inspection approach described above (Kleinberg *et al*, 2007). The reason anthropometry fails is that the small metric differences between faces are easily swamped by changes in lighting, pose, expression, and even the focal length of the camera lens. Simple metrics do not survive such variability, so images of different faces can easily give rise to more similar signatures than images of the same face.

In view of the difficulty of matching unfamiliar faces, it is significant that photo-ID documents continue to be central to national security policies. In the UK, the introduction of a national ID card that includes a photograph of the holder is expected to begin in 2009, following consolidation of the Identity Cards Act 2006. 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 to the face of the traveller (eg, the SmartGate system under development by Australian Customs at Sydney airport). For several reasons however, the advent of machine systems does not represent a solution to face recognition. First, the performance of automatic face recognition systems in the past does not lead one to be optimistic about their imminent ability to solve the problem. Even under highly restricted conditions, and dealing with cooperative individuals, today's best systems are far from infallible. When conditions are not so controlled (as with surveillance CCTV), or cooperation is poor (for example, when someone is trying to conceal his or her identity), the accuracy of these systems plummets. Secondly, these systems are designed to minimize "false negative" errors, ie, to ensure that fraudulent documents are not accepted as legitimate. This means that they tend to generate high numbers of "false positive" errors, in which legitimate documents are challenged. The task of rechecking all the rejected photo-ID documents then falls to the human operators who are required to make a final decision in each case. Thirdly, the development of technological solutions addresses only a rather limited number of situations in which photographic identification is used. While some security checks in some airports may employ the best available automatic face recognition systems, routine identification decisions are required in many less well-controlled settings (from surveillance CCTV to authorizing a customer's payment in a shop).

The basic message is that neither humans nor machines can reliably match unfamiliar faces. How does this relate to our experience of recognizing faces in everyday life? Recent

psychology research indicates that familiarity is the key. As faces become increasingly familiar, the observer's ability to match them improves dramatically. In an early demonstration of this, Burton *et al* (1999) found that students could match two images of their lecturers almost perfectly, even when one of the images was taken from very low-quality CCTV footage. Using exactly the same images, viewers unfamiliar with the people performed at chance levels on this test. Furthermore, police officers were no better at this task than anyone else who did not know the people in the pictures. This study, and other work conducted over the past 10 years, has demonstrated that a key predictor of one's ability to match faces is one's level of familiarity with the person depicted (eg, Clutterbuck & Johnston 2004, 2005). So, the task of validating photo-ID documents is difficult because it usually requires operators to match two images (normally one photographic and one corporeal) of an unfamiliar person: exactly the conditions which experimental work has shown to be so hard.

As these studies indicate, familiarity is the key to successful face recognition. In several recent studies, we have found that simulating familiarity for automatic face recognition can boost performance in machine-based systems as well (Burton *et al*, 2005; Jenkins *et al*, 2006). The principle is straightforward. In the case of human performance, familiarity seems to be a natural consequence of increased exposure. To model increased exposure, we collected several different photographs of each person, and averaged them together to create a single image for each face. The resulting images are quite uncanny, and seem to bring out the true essence of each face (see <http://www.psy.gla.ac.uk/~mike/averages.html>). For example, although different photographs of Bill Clinton are rather varied images, when averaged together they form a striking Clinton likeness that is recognized very well by humans and automatic face recognition systems alike. In fact, in our studies, these identity averages were recognized better than the photographs from which they were constructed. This is because the averaging process washes out aspects of the image that are unhelpful (such as directional lighting), while consolidating aspects of the image that are diagnostic of identity (specifically the physical structure of the face). The average image of a person's face has several intriguing properties that are highly desirable from an identification standpoint. First, the average stabilizes surprisingly quickly. Even when only two photographs of the face are available, performance is much 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 – the average of any set looks 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 these systems are not required to work from the same source images. Thirdly, the averaging process is extremely robust against errors. Errors must be considered inevitable in a large database, so it is essential that the system

does not collapse when errors arise. Our studies show that an average image is highly resistant to contamination from misidentified photographs. As long as the average is constructed from enough images (16 or so), recognition is barely affected, even if two or three of the photos come from a different person.

The breakthrough of better-than-photo recognition accuracy raises the question of whether face databases and ID documents should contain stabilized average images, rather than standard photographs. One pragmatic advantage of this proposal is that it only involves 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. Note that our proposal does not seek to supersede previous advances in automatic face recognition research. Instead, it makes an independent contribution that can plug into existing infrastructure for added benefit. We conclude that although unfamiliar face matching is poor in humans and machines alike, averaging together different photos of the same person can improve performance enormously. This striking result has important implications for the future of automatic face recognition, as well as for crime prevention and national security policies. Our findings also demonstrate that with face recognition, as with so many other problems, we can improve machine performance by mimicking nature's solution.

References

- Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., 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., Wilson, S., Cowan, M. & Bruce, V. (1999). "Face recognition in poor quality video: evidence from security surveillance". *Psychological Science*, 10, 243-248.
- Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). "Robust representations for face recognition: The power of averages". *Cognitive Psychology*, 51, 256-284.
- Clutterbuck, R. & Johnston, R. A. (2004). Matching as an index of face familiarity. *Visual Cognition*, 11, 857-869.
- Clutterbuck, R. & Johnston, R. A. (2005). "Demonstrating how unfamiliar faces become familiar using a face matching task". *European Journal of Cognitive Psychology*, 17, 97-116.
- Jenkins, R., Burton, A. M., & White, D. (2006) "Face recognition from unconstrained images: Progress with prototypes". *Proceedings of the Seventh IEEE International Conference on Automatic Face and Gesture Recognition*, 25-30.
- Kemp, R., Towell, N., & Pike, G. (1997). "When seeing should not be believing: Photographs, credit cards and fraud". *Applied Cognitive Psychology*, 11, 211-222.
- Kleinberg, K. F., Vanezis, P., & Burton, A. M. (2007). "Failure of anthropometry as a facial identification technique using high quality photographs". *Journal of Forensic Science*, 52, 779-783.
- Megreya, A. M. & Burton, A. M. (2006). "Unfamiliar faces are not faces: Evidence from a matching task". *Memory & Cognition*, 34, 865 – 876.
- Megreya, A. M. & Burton, A. M. (2007). "Hits and false positives in face matching: A familiarity-based dissociation". *Perception and Psychophysics*, 69, 1175-1184.
- Murakami Wood, D. (ed.), K. Ball, S. Graham, D. Lyon, C. Norris and C. Raab. (2006). A report on the surveillance society. Office of the Information Commissioner. Wilmslow, UK.
- Porter, G., & Doran, G. (2000). "An anatomical and photographic technique for forensic facial identification". *Forensic Science International*, 114, 97-105.
- 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". *Journal of Forensic Science*, 41, 786-791.
- Wells, G. L. (1993). "What do we know about eyewitness identification?" *American Psychologist*. 48, 553-571.
- Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). "From the lab to the police station: A successful application of eyewitness research". *American Psychologist*, 55, 581-598.

DRIVING OFFENCES RESULTING IN DEATH

The Sentencing Guidelines Council has published a consultation that covers four offences; causing death by dangerous driving, causing death by careless driving under the influence of alcohol or drugs, causing death by careless driving and causing death by driving: unlicensed, disqualified or uninsured drivers.

It recommends that prolonged, persistent and deliberate bad driving and consumption of substantial amounts of drugs or alcohol should put offenders into the most serious category of causing death by dangerous driving.

A combination of these features of dangerous driving – particularly if accompanied by aggravating factors, failing to stop or a very bad driving record – should attract sentences towards the maximum term of 14 years.

For most of the offences, Judges and magistrates will need to assess how bad the driving was and the degree of danger

that it created. Other issues – largely related to the offender's behaviour – are treated as aggravating factors.

The Council advises a robust approach to the use of mobile phones; the fact that an offender was avoidably distracted by a hand-held mobile phone when the offence was committed will always make an offence more serious, the guideline says.

In dealing with cases of causing death by careless driving under the influence of alcohol or drugs, the proposed guideline gives greater weight to the degree of intoxication of offenders. Sentences for cases of careless driving just falling short of dangerous driving should have a starting point of 15 months' imprisonment, the Council recommends.

In cases involving "momentary inattention" and no aggravating factors, offenders should be given a community sentence which could include a curfew requirement. The guideline is available at: www.sentencing-guidelines.gov.uk. The consultation closes on March 10, 2008.