CITATION: White, D. & Kemp, R. (in press). Identifying people from images. In *Improving the criminal justice system: Perspectives from psychological science*, Brewer, N & Bradfield Douglas, A. (Eds.), Guilford Publications.

Identifying people from images

David White & Richard Kemp
UNSW Sydney, Australia

Forthcoming in *Improving the criminal justice system: Perspectives from psychological science*
(Eds. N. Brewer, A. Bradfield-Douglass)

Identifying unfamiliar people from images is critical to crime prevention, criminal investigation and in court deliberations. Is the applicant who they claim to be? Does the suspect match the culprit captured on CCTV? Is the person depicted on this wanted poster someone I know? The criminal justice system relies on the accuracy of these decisions, but is this reliance warranted?

In this chapter we review psychological studies examining the many ways that images are used to identify people in these settings. We start with the task of verifying identity from photo-ID. Establishing a person's identity is key to crime prevention, criminal investigation and identity fraud, which can be the precursor to serious and organized crime. Because images on Photo-IDs – for example driver licenses or passports – are typically subject to strict quality control measures, this also represents optimal conditions for matching faces of unfamiliar people. Therefore, studies examining accuracy on this task provide a useful baseline of human accuracy in face identification, and for understanding problems that arise when using images for identification in court.

We then turn to the more challenging case of identifying people from CCTV. This is the most common image identification evidence presented in court and is becoming a particularly important research area in light of increased levels of surveillance in modern society. In the final section, we review literature evaluating methods for generating image likenesses of culprits from a witness' memory.

PHOTO IDENTIFICATION

Photo-ID remains the most common method for verifying a person's identity. We rely on images of faces to link cardholders to biographical details on their identity documents, and as a result, place substantial trust in the ability of passport officers, police and security professionals to decide whether or not the unfamiliar face of the cardholder matches the photo on their identity document. But how accurate are these decisions? People often assume that the task is trivial. First, the task does not involve memory – the image is compared to the person standing in front of you. Second, passport images are subject to strict guidelines to ensure they are high quality, taken in standardized conditions and under good lighting. Third, we routinely recognize people in our daily lives from the briefest glimpse of their face.

Perhaps for these reasons, people are often surprised to learn that performance on these face identification tasks is highly error-prone. In an early study, Kemp and colleagues (1997) asked supermarket cashiers to verify the identity of participants posing as shoppers, by comparing their appearance to a photo-ID card. Shoppers presented their photo-IDs to cashiers, who then decided whether the photo matched the card bearer. On half of the trials, IDs were 'valid', meaning that the photo was an image of the shopper taken in the weeks prior to the experiment. In the other half, 'invalid' images of another person were presented. Invalid IDs that were chosen to resemble the shopper were incorrectly accepted by the cashier on over half of the trials, and overall error on the task was 35% -- not much better than accuracy that would be expected by random guessing.

In the late 1990s a number of lab-based studies replicated Kemp et al's finding by examining face matching performance under more optimal conditions. For example, in a series of studies by Vicki Bruce, Mike Burton and colleagues (e.g. Bruce et al. 1999, Henderson, Bruce & Burton, 2001), participants were provided with a target face above an array of ten images that may or may not contain the target identity. All were high quality images, taken on the same day, in the same neutral pose and under very similar lighting conditions. This computer-based task was designed to provide an analog to police line-ups -- with the important difference that the task does not involve memory. An example of the face matching task is shown in Figure 1.

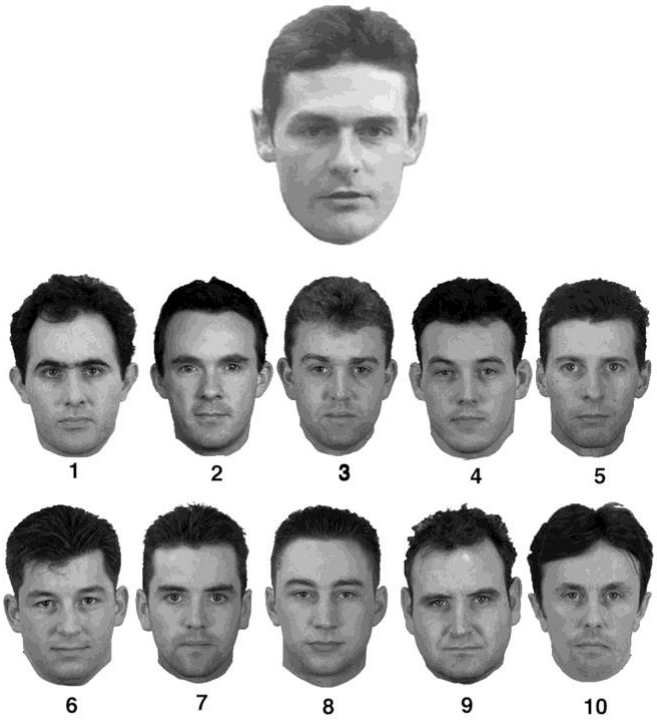

Figure 1. *A typical example of a trial from a one-to-many face matching study showing the level of similarity between the faces in these studies (Bruce et al., 1999; see also White, Dunn, Schmid, & Kemp, 2015). Participants must decide if the target person pictured above the array appears in the array, and if so to decide which image matches the target. The correct answer is shown in the Author Note at the end of this chapter.*

Despite these favorable conditions for matching, participants in Bruce et al.'s (1999) experiments made errors in approximately 30% of decisions. Subsequent studies have replicated this poor level of accuracy under a variety of conditions. For example, replacing the target photograph with a live person does not improve accuracy (Megreya & Burton, 2008). Other studies have reduced task demands further, by presenting two images side-by-side on a computer monitor, and asking participants to decide if they are of the same person or two different people. This does not redress task difficulty, with error rates in pairwise matching decisions typically around 20% (Henderson, Bruce & Burton, 2001; Megreya & Burton, 2006, 2007; Burton, White & McNeill, 2010).

Given the reliance that society places on these decisions, face-matching errors in professional roles can carry potentially serious outcomes. But does professional

experience protect people from making these errors? Early work by Burton et al. (1999) asked police officers to match images of unfamiliar faces with CCTV footage. Despite being more confident in their identification decisions, police officers were not any more accurate. While police officers' apparent obliviousness to the difficulty of the task is concerning, face matching was not explicitly part of officers' job descriptions. More recently however, White et al. (2014) tested the performance of Australian passport officers who are explicitly required to match faces routinely in their daily work. Surprisingly, despite receiving training in face identification, passport officers were no better at face matching than a group of novice university students.

Why is unfamiliar face matching so difficult?

Unfamiliar face matching tasks do not require participants to memorize faces. Images are presented on the screen simultaneously, and participants typically can take as much time as they like before reaching a decision. So, the difficulty of this task is not caused by the fallibility of human memory (see Brewer, Sauer & Palmer, this volume), but appears to be a perceptual limitation. Despite the best efforts of passport issuing authorities to optimize the quality of passport photos – ensuring for example that they are evenly lit, facing the camera and showing neutral pose – the evidence we have reviewed above clearly shows that people make large numbers of identity verification errors. So, why is this perceptual task so difficult?

Figure 2*. Top row: three images of the same individual taken seconds apart, but from different distances (from left to right: 50cm, 100cm, 300 cm); from Burton, Schweinberger, Jenkins & Kaufmann, 2015. Bottom rows: Passport-compliant photographs of 2 people (rows), all taken on the same day but by different passport photo vendors (columns); from Spiteri, Porter & Kemp, 2015.*

At least part of this difficulty can be explained by the intrinsic limitations of photography in capturing facial identity. For now, let's ignore the changes in facial appearance caused by aging, expression, head angle and lighting, and focus only on the optimal conditions for matching where all of these variables are controlled. Figure 2 shows three people all pictured on the same day and under controlled conditions. The top row shows the same individual taken just seconds apart, in precisely the same studio conditions and with precisely the same camera. Nevertheless, the change in appearance from the leftmost image to the rightmost image is striking, and is caused simply by the person placing himself further away from the camera. This simple change in subject-to-camera distance has plainly altered the perceived shape of this person's face (Burton et al., 2015). In a recent study, Noyes and Jenkins (2017) examined the effect that this change has on accuracy, while controlling for all other variables. This simple change has a substantial impact on the accuracy of face matching decisions, reducing accuracy on a same or different person decision by 10%.

Now consider the bottom two rows of Figure 2, which show five passport compliant photographs of two individuals. In each column, images are from a different passport photo vendor in the same local area. Despite these images all being taken on the same day, in neutral pose and conforming to passport image guidelines regarding lighting and head angle, they nevertheless give rise to quite different appearances. This is partly due to the different lens and sensor characteristics of the cameras, which has a rather marked effect on the appearance of skin tone, hair colour and face shape. This example underlines the essential difficulty of identifying people from photographs: no matter how hard one tries, it is very difficult to ensure that the same face appears the same in any two photographs.

Naturally, optimal conditions for matching are rarely encountered outside of the laboratory. So far, we have presented accuracy scores for the most straightforward matching tasks. However face matching accuracy is reduced further by a range of viewer- and face-related factors that are encountered in the real-world, including: aging of the face (Megreya, Sandford & Buron, 2013); disguise (Noyes & Jenkins, 2016); changes in lighting, pose and expression (Hancock, Bruce, & Burton, 2000; Jenkins, White, Van Montfort, & Burton, 2011); time pressure (Fysh & Bindemann, 2017); lack of sleep (Beattie et al. 2016) and state anxiety (Attwood, Penton-Voak, Burton, & Munafò, 2013).

Why do people assume unfamiliar face matching is easy?

Participants are often surprised at the difficulty of matching images of unfamiliar faces, and predict it will be a straightforward task. This may explain why people have been relying on Photo-ID ever since photography made the practice possible (Bertillon, 1893), and yet it is only relatively recently – in the past twenty years – that scientists have discovered the practice is largely ineffective. Perhaps instead of asking *why this task is so difficult*, we should instead be asking – *why do people expect it to be easy*? Where does our intuition that Photo-ID is a reliable method for identity verification stem from?

Figure 3*. Two face matching decisions – do these pairs of images show the same person or different people? The image pair on the left is an item from the Glasgow Face Matching Test (Burton, White & McNeill, 2010) and the pair on the right shows a familiar person. Answers are provided in text below.*

A recent proposal is that we overgeneralize our expertise in recognising *familiar* faces to the case of unfamiliar face matching (see Jenkins & Burton, 2011). We are very good at recognizing faces of people we know, and we experience the effortless recognition of these faces many times each day. Perhaps we think faces are useful tokens of identity because we recognize familiar faces so effortlessly? As can be seen from Figure 3, familiarity transforms face matching tasks from a simple image-to-image comparison to a task of *recognition*. The image pair on the left shows an item from the Glasgow Face Matching Task (Burton et al. 2010). This is a difficult item, and around of a third of people incorrectly answer that they believe the images are of two different people. In the righthand image pair, there are substantial disparities in age, pose, expression, image quality, make-up and distance-from-camera. However, most people have no difficulty in recognizing this person, and hence deciding that these images are of the same person.

Ritchie et al. (2015) have recently provided some empirical support for the hypothesis that people overgeneralize expertise with familiar faces. They asked participants to complete pairwise face matching decisions like those presented in Figure 3. Participants had to first decide if the two images were of the same person or of different people. Half of the image pairs were of local UK celebrities and so were familiar to the UK participants tested in this study, and half were local Australian celebrities that were unfamiliar to the UK participants. Critically, participants also had to estimate the proportion of German participants – i.e. people who were unfamiliar with all the faces in the study – who would get this decision right. Consistent with other work, face matching accuracy was far better for familiar than for unfamiliar celebrities, but most interestingly, UK participants also predicted that the German participants would perform better on UK than Australian celebrity pairs – the faces with which they themselves were familiar.

This result may go some way to explain why people tend to be overconfident when performing unfamiliar face matching tasks (see Bruce et al. 1999; Burton et al. 1999). Our misplaced reliance on our ability to identify *unfamiliar* faces may stem from intuitions based on our experience with *familiar* faces. Regardless of what is causing overconfidence, this misconception may help explain why we are not more aware of our inability to match unfamiliar faces. Indeed, this overconfidence may itself be at the root of the problem we have outlined at this section. Poor performance in

unfamiliar face identification tasks is necessarily a problem if people are aware of it: With awareness one can take steps to mitigate the problem. The greatest danger arises when people are both confident and wrong, and so it will be important for future research to address why this misplaced confidence arises and how it can be redressed.

FORENSIC FACE IDENTIFICATION

Cameras are everywhere in the modern world. Although not unique in this regard, British people appear to have a particular obsession with recording themselves, and it is a belief often held that the UK is the most closely observed society in the world. One recent estimate suggests that in 2016 there were around 5 million CCTV cameras in the UK, a nation with a population of about 65million (British Security Industry Associations, 2015). This estimate includes cameras facing public spaces such as streets and parks and also cameras monitoring private places such as shops, schools, work sites, store rooms etc. Critically, this estimate doesn't include the cameras most adults carry with them in their smartphone. As a result of the ubiquity of recording devices, today it is rare for media coverage of any significant news event not to be accompanied by a montage of video clips from public and private fixed cameras and hand-held smartphones.

The fact that the world is now so closely monitored impacts our lives in many ways, including on the operation of the legal system. It is now routine for offenders to be recorded while committing an offence, and these recordings can be critical to police investigation and may become evidence in cases ranging from vandalism and motoring offences through to robbery, murder and terrorist attacks. In such cases, a critical issue is the identity of the culprit depicted in the CCTV images. Broadly, there are two distinct processes likely to operate here:

1. Investigators who have no suspect in mind may use the surveillance image of the culprit to search databases of potential suspects. This is the modern day equivalent of asking a witness to search through "mugshot" books.

2. Once a suspect has been identified and charged, prosecutors need to present evidence that the offender depicted in the images is the person charged with the offence. In some court cases this becomes the central legal question; both prosecution and defense may accept most of the facts of the case leaving only the identity of the offender in question.

One-to-many searches using surveillance images

In the first of these scenarios, surveillance images are used as a search template in what is called a "one-to-many" search of a database. One surveillance image is compared to each of the images in the database using face recognition software. However it is important to note that this software does not identify the suspect in a database, rather it ranks the database in terms of apparent similarity to the search template and returns a 'candidate list' of the highest ranked images to a human operator to consider. Well-designed algorithms, trained on appropriate datasets, and searching in relatively small databases of good quality images will often, but not always, return a ranked list which contains the target in the first or second rank (Grother & Ngan, 2014).

However, this is not inevitably the case. The target may not be present in the database, in which case the highest rank return will be of a very similar looking person who is not the target. Even if the target is present in the database, they may appear far down in the ranking – perhaps due to image quality issues, or alterations in appearance resulting from ageing or changes to facial paraphernalia, including eyeglasses and hair. The issue of image quality is especially critical when using surveillance images from CCTV to search databases. For operational reasons and to prevent tampering, CCTV cameras are normally positioned several meters off the ground on buildings and poles, leading to a difficult angle of view, and these systems are often designed to cover large areas, resulting in distorted, low resolution images of the offender. Further data loss may occur if the images are compressed for storage, and many CCTV systems only save one or two images per second (Edmond et al, 2009).

If we add to this the fact that there may be variations in lighting leading to shadows and over exposed areas in an image, that many offences occur at night and that offenders often wear head coverings such as peaked caps and other devices deigned to mask the face, then it becomes apparent that surveillance images are commonly of very poor quality, showing low resolution, partial views of offenders captured from difficult angles of view. Recent benchmarking tests of leading face recognition software shows that these face recognition algorithms are particularly error-prone when comparing images captured in these conditions (Phillips, Hill, Swindle & O'Toole, 2015).

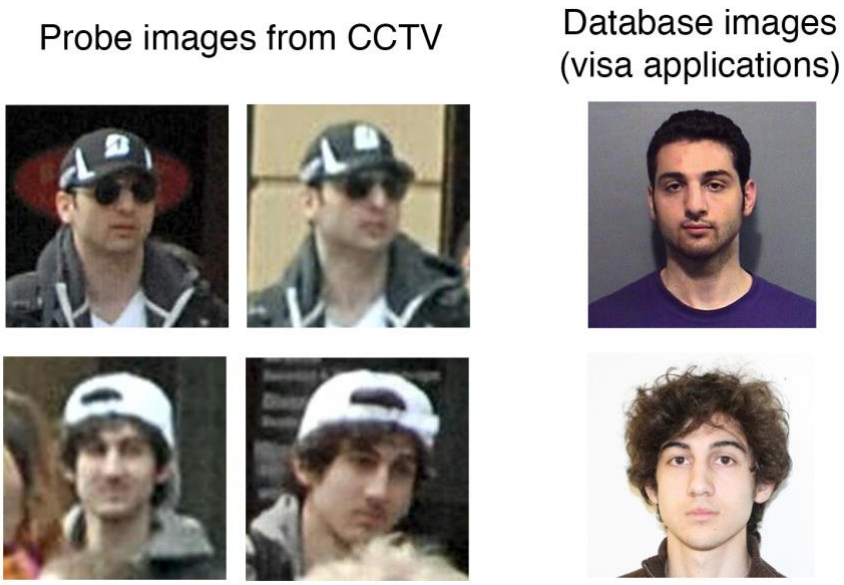

Figure 4*. CCTV images of the 'Boston Bombers' released by the FBI (left). We now know these images were used to search image databases containing high quality images of the brother (right),* but that these searches failed to identify the terrorists who were ultimately recognized by a relative.

The limitations of using face recognition software to identify suspects was underlined in the search for the perpetrators of the Boston marathon bombing in April 2013. Police quickly located relatively good quality surveillance images of two suspects who were nicknamed "white hat" and "black hat". These images, shown in Figure 4, were released to the public soon after. The *Washington Post* reported that, in

the hours following the bombings, these images were used to search several databases which, we now know, contained driver's licenses and other images of the bombers (Montgomery, et al 2013; Klontz & Jain, 2013). However, face recognition software failed to identify the suspects. Instead, the aunt of the two brothers pictured on CCTV recognized them and reported their identities to the FBI.

Although not without limitations, it is clear that one-to-many searches of databases using surveillance images are becoming an important feature of crime investigations. The US and Australia now have nationwide systems that enable police officers to perform one-to-many searches of citizenship, mugshot and driver's license databases (Garvie, Bedoya, & Frankle, 2016). Importantly, human adjudication is required to examine the possible matches returned by the computer system. Coincidently, this task is very similar to the Bruce et al. (1999) line-up task that is illustrated in Figure 1. Given what we know about human performance in this task, this is a potentially dangerous situation. A recent study by White et al. (2015) showed that passport officers who use this software in their daily work make errors on 1 in every 2 candidate lists they review, despite all images in these tests being high-quality and complying with passport standards. Even more concerning is that in 40% of trials, these passport officers selected a person that was not the target as a match.

So, while these systems provide a new weapon in the fight against crime, they also have substantial potential to waste valuable police time following up false leads. More seriously, in searches of databases containing plausible suspects – for example, past offenders – this process poses a significant risk of wrongful convictions in the future. Moreover, the accuracy of face recognition software can be biased towards making errors when searching for ethnic minorities (Phillips et al., 2011). Combined with the fact that humans are also more likely to make errors when identifying faces from different ethnic groups to their own (see Meissner & Bingham, 2001; Megreya, White, & Burton, 2011), this raises the additional concern that face identification systems will build racial bias into the criminal justice system (Garvie, Bedoya, & Frankle, 2016).

Increased use of face recognition software in criminal investigation has not been accompanied by improved understanding of their operational accuracy. Operational accuracy relies on a complex interaction between computer performance, human performance and properties of the images being searched (see Towler, Kemp & White, 2017; White, Norrell, Phillips & O'Toole, 2017), and so it will be important for researchers in this area to adopt an interdisciplinary outlook in the future. Such an approach is necessary to provide accurate estimates of system performance and to design ways in which to improve system accuracy. Notwithstanding, psychological research can play a key role in this emerging field, both in understanding human performance in computer-assisted face identification tasks and in understanding how people perceive, understand and interact with this technology.

Ultimately, a critical safeguard against false convictions stemming from facial image evidence will be the legal processes that occur at trial. Prosecutors will be required to convince the triers of fact – judges and jurors – beyond a reasonable doubt that the defendant is the person shown in the surveillance images. But how accurately will the courts be able to make this one-to-one matching decision? We turn to this question in the next section.

One-to-one matching of surveillance images

The way in which courts make use of surveillance images has changed over time and varies between jurisdictions (Edmond et al., 2009, Edmond et al., 2010). Some courts leave it to the jurors to determine whether the defendant is the person depicted, while other courts have allowed forensic image analysts to give expert evidence to help them interpret the images. We now consider the evidence regarding the ability of these two groups to make these identification decisions.

Can jurors identify defendants from surveillance images?

The clear conclusion from studies of unfamiliar face matching conducted over the last two decades, which we summarized earlier in the chapter, is that this is a difficult and error-prone task. This difficulty is confounded by the fact that people tend to overestimate their ability to determine whether two images are of the same person (e.g., Burton et al., 1999; Ritchie et al., 2015). In the forensic setting, this makes it likely that jurors will make false positive errors, incorrectly concluding that a surveillance image is of the defendant, and their judgement may be further influenced by the suggestive context of the courtroom and the fact that no alternative suspects are offered. In almost all cases, jurors have to make a one-to-one matching decision without having seen any alternative candidates.

A few studies have sought to model the situation faced by jurors in such cases. Across three experiments Davis and Valentine (2009) tested participants' ability to determine whether the culprit seen in a video clip was the defendant who stood in front of them. Experiment 1 employed eight different defendants and video clips showing the culprit both in full-face and profile, while occupying at least half of the frame. Even given these unrealistically good video images, performance was far from perfect. Participants incorrectly identified the defendant in 17% of cases where the video showed someone else (i.e., they made a false positive error akin to convicting an innocent defendant), and on 22% of cases failed to identify a "guilty" defendant who appeared in the video. Importantly, accuracy varied across defendants. For example, one defendant was always identified when present in the video (100% hit rate) but was also likely to be falsely identified when innocent (44% false positive rate). In contrast, another defendant escaped conviction on 36% of occasions and was falsely convicted on just 5% of trials.

Thus, not only are the error rates high overall, they also vary greatly – probably due in part to the degree of resemblance between the defendant and the individual acting as the similar looking culprit. Experiment 3 in this series was designed to investigate whether the identification errors seen in Experiment 1 were also the result of limitations in the quality of the surveillance video. In this study the surveillance footage was replaced with high quality videos showing the culprit's face in frontal and profile views. Over a quarter (26%) of participants failed to convict a guilty defendant who stood in front of them as they watched a video recorded just a week earlier. Even when the video was only 1 hour old, participants failed to identity the guilty defendant from the video in 17% of cases. More worrying still, participants wrongly identified an innocent defendant in over 40% of cases.

Overall, this evidence suggest that judges and jurors are likely to be prone to making errors, including false identification errors, when asked to determine whether the defendant is the person seen in a surveillance image. In courts that rely on expert witnesses to help interpret these images, do these experts fare any better?

Can "experts" identify defendants from surveillance images?

Courts in several jurisdictions have grappled with the issue of how to deal with identification evidence from surveillance images (Edmond et al., 2009). For example, in Australia prosecutors initially attempted to use evidence from police officers who claimed to be able to identify the defendant in the images tendered as evidence. This approach was rejected by the courts on the basis that the police officers had no training or expertise that would allow them to make more accurate identification decisions than the jury. In response to these rulings, prosecutors sought out expert witnesses who could analyse the surveillance images, leading to a series of cases in which identification evidence was presented and supported by expert evidence from Forensic Image Analysts, or as the press sometimes termed these individuals "Facial mappers" (Edmond et al., 2009). A similar process has occurred in other countries, leading to the emergence of groups of individuals claiming this specialist identification expertise.

Cases have come to light in which expert evidence has proven to be wrong. The first of these cases involves expert evidence provided in a murder case in the UK, as reported by barrister Campbell-Tiech (2005). In this case police asked four different facial mapping analysts to compare photographs of the suspect to surveillance images, and all four agreed that there was some support for a match. Sometime later, the investigators decided that they had arrested the wrong person and named a new suspect, whose image was sent to these same four analysts. Of the four, the first two now reached "inconclusive" findings, the third said there was support for the conclusion that the surveillance images did not depict the new suspect and the fourth concluded that there was "powerful support" for the conclusion that they were the same person. Presumably troubled by the third expert's conclusions, the police asked this person to reconsider their evidence, making it apparent that they believed the new suspect was the person shown on the CCTV. This expert now reported that he could not exclude the possibility that it was the same person. Thus this one piece of surveillance video had been linked to two different suspects with widely varying levels of identification confidence.

The second case of a known ID error is even more bizarre. In 2009 the Australian newspaper *The Sunday Telegraph* published 30 year old pictures of a semi-naked woman who they wrongly claimed was politician Pauline Hanson. Ms. Hanson denied the images were of her, and a few days later after other facts emerged, the newspaper apologized and retracted the story (Breen, 2009). However, shortly before this retraction, several forensic image analysts were asked their opinion of the images, including two individuals who at the time were regularly giving facial mapping evidence in Australian courts, Professor Maciej Henneberg and Dr Meiya Sutisno. While Dr Sutisno concluded that the images were probably not of Ms. Hanson, Professor Henneberg, was reported as saying that the photographs were "99.2 per cent sure" to be of Ms. Hanson after apparently calculating that there was a 0.8% chance that two people would share such similar features (Leys, 2009). Thus, two experienced experts who regularly testify in criminal matters gave diametrically

different opinions when asked to compare the high quality photos to images of Ms. Hanson.

In both these cases an individual who claims expertise in forensic image analysis, and whose evidence has been accepted in court, has been shown to have made an identification error. In the Hanson case this error occurred even though the images under consideration were high quality (we do not have permission to reproduce the images here but inquisitive readers can easily find them on the web if so inclined). Of course, errors made by this small sample of 'experts' may not be reflective of accuracy in this profession more broadly and methods used by forensic facial examiners vary from one examiner to the next. Nevertheless, some key approaches used by experts in court have been shown to be unreliable, such as the practice of identification by measuring distances between facial features, known as 'facial mapping' (e.g., Kleinberg, Vanezis, & Burton, 2007), and the use of certain digital tools (Strathie, McNeill, & White, 2012; Strathie & McNeill, 2016). This has resulted, in recent years, to the creation of international standards for facial comparison practitioners (Facial Identification Scientific Working Group, 2012).

More recently, psychologists and forensic scientists have begun to conduct systematic tests of expert accuracy in facial image comparison. Initial results from these studies show that experienced facial comparison experts do outperform novices. For example, Norell et al. (2015) compared the performance of a group of 17 forensic facial analysis experts to untrained students. Participants compared a high quality reference photograph captured some months or years earlier to a "questioned image" and indicated their response and confidence, using a nine-point scale similar to that used by many practitioners. The questioned image was either high, medium or low quality and was designed to approximate a surveillance image. Overall, the expert made slightly more correct decisions (76% vs about 72%) and fewer errors (3% vs about 21%) than the novices. The experts were more likely to make use of the inconclusive mid-point of the scale, and this was especially true when examining lower quality images. However, it is important to note that the experts made several errors and were only error free when examining the highest quality images showing the same person. When the images were of different people, even the high quality photographs resulted in some erroneous positive identification decisions by the experts. Given that many of these experts "apply their knowledge in casework for legal authorities", these errors must cause some concern.

The best evidence that expert groups can outperform novices comes from a study which tested the performance of a group of forensic facial examiners attending an international facial biometrics meeting organized by the FBI (White, Phillips, Hahn, Hill, & O'Toole, 2015). This group were compared to university students and a control group of non-facial examiners who attended the meeting in another professional capacity – for example, because they administered biometric systems or performed managerial roles. All three groups were tested on the same battery of tests of facial perception, including the Glasgow Face Matching Test (GFMT; see Figure 3), and two other challenging tests designed for this study. Across all tests, the examiners outperformed the control group, who outperformed the student group.

In one of the new tests designed for this study, examiners outperformed both other groups, particularly when given longer to make their decisions (30 seconds vs. 2

seconds). This is the best evidence we have for superior face matching performance by "expert" groups, but again it should be noted that experts did make errors. However, this study may not provide a definitive picture of accuracy in these expert groups. First, the methods used in this study did not permit experts unlimited time to reach their decisions. Second, experts were not provided access to digital tools or procedural documents that would usually support their decision making. Third, the images used in these tests were of relatively good quality and were not reflective of surveillance imagery that would typically be analysed in casework (for example, all faces were looking straight at the camera, which is very rare in CCTV images). It remains to be seen how professional examiners perform under these conditions.

Overall, the current experimental evidence suggests that some groups of forensic image analysts are likely to be more accurate than the general population (for a detailed review see White, Norell, Phillips, & O'Toole, 2017). However, even with good quality images that are not representative of typical casework, these experts make identification errors, including false positive identification errors that could potentially lead to the imprisonment of innocent suspects.

Does practice make perfect?

Why do these expert groups outperform novices? This is a critical question from a legal perspective, because judges, lawyers and jurors must decide whether claims to expertise by facial comparison analysts are founded on legitimate grounds (see Cutler, Marion & Kaplan, Chapter 14).

For example, many experts claim that their expertise rests on their professional experience in this area. One might reasonably expect that the superior abilities of facial forensic examiners is the result of many years of practice at matching unfamiliar faces. Although plausible, current evidence suggests this is not the case. White, Kemp, Jenkins, Matheson, and Burton (2014) examined the performance of a group of passport officers with up to 20 years of experience who, as part of their daily work to validate passport applications, were required to make photo-to-photo and photo-to-person comparisons. In the first test they were required to decide whether a photograph presented on a computer screen matched the person standing in front of them. The passport officers falsely accepted 14% of the fraudulent applications presented, and falsely rejected 6% of valid photographs. In a second test, passport officers completed a photo-to-photo comparison test in which they were required to match recent photographs to images captured two years earlier. Participants made errors on about 30% of match trials and about 11% of mismatch trials. Critically, the passport officers were no more accurate than the inexperienced students, and the same result emerged when the two groups were compared on a standardized test of face matching, the Glasgow Face Matching Test.

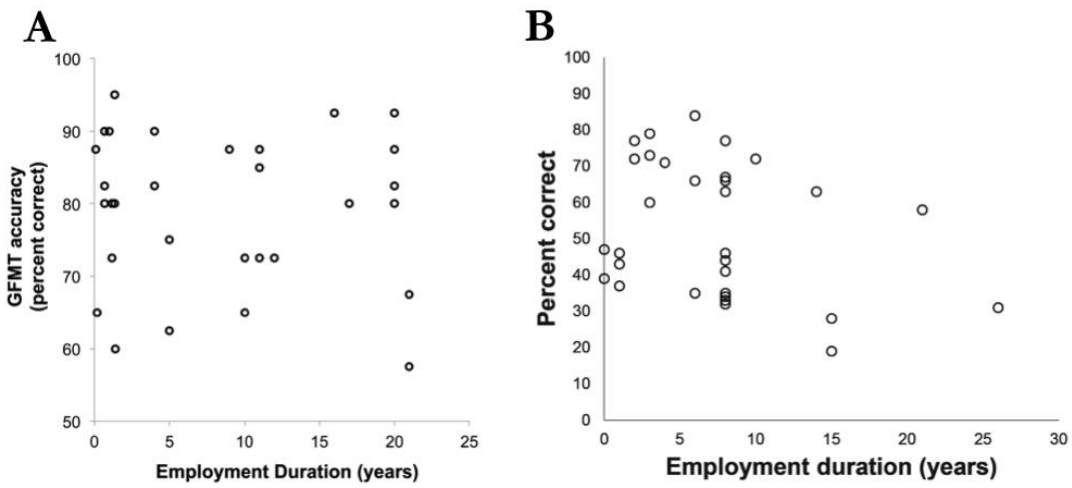

Figure 5*. Relationship between length of service as a passport officer and performance on face matching tasks (A: White et al. 2014, B: White et al. 2015).*

Perhaps the most compelling finding was that, across all three tests, there was no association between length of employment experience and accuracy: some new employees achieved around 95% accuracy on the GFMT while others with 20 years of experience performed little better than chance (see Figure 5). This same pattern has been observed in a more recent test of German border control officers. Indeed, Wirth and Carbon (2017) report that performance was actually *worse* in individuals with longer service. Moreover, across both studies, the vast majority of errors were made when falsely accepting two non-matching photographs as showing the same person – precisely the type of error that security professionals should be aiming to avoid! Apparently then, performing unfamiliar face tasks repeatedly in daily work is not sufficient to improve performance. This finding has important implications for both recruitment of staff that are required to identify faces, and also when assessing expertise in court.

Can face identification be trained?

If experience alone cannot account for the superior performance of some expert groups, then perhaps the answer lies with the training these individuals have received. Many police and government bodies around the world have developed training programs for individuals that perform unfamiliar face matching tasks in their daily work. In a study employing low quality video clips that are typical of CCTV footage, Lee, Wilkinson, Memon, and Houston (2009) investigated whether individuals with training in forensic facial identification made accurate identification decisions. A small group of fifteen graduates from an MSc degree in Human Identification with varying amounts of experience were compared to a group of untrained participants. However, graduates and untrained groups had similar error rates, making correct identification decisions in only about 67% of cases, and false positive errors in about 22% of cases. Thus in around a quarter of cases, these experts wrongly identified the defendant as the person in the surveillance images. Furthermore, experience had little impact; graduates with three or more years of professional experience in the field were no more accurate than those with up to one year of experience.

More recently, Towler (2017) undertook an analysis of professional facial comparison training programs and identified a number of common components, including training in facial anatomy and forensic photography. To test the impact of training in these domains the researchers asked students to complete a standardized test of face matching performance before and after they undertook relevant University level courses, and found no evidence of any improvement following training. This result is consistent with other before-and-after evaluations of workplace training in face identification (Woodhead, Baddeley, & Simmonds, 1979). In addition, laboratory studies show that certain strategies that are taught in these training courses are ineffective (Towler, White, & Kemp, 2014).

However, one other component of these training programs does appear to have merit. Many training programs promote a feature comparison technique, which encourages practitioners to compare the face feature-by-feature rather than in a more holistic manner. Interestingly, this approach contrasts with the evidence from studies of familiar face recognition that accurate recognition of familiar faces is supported by holistic rather than feature-based processing (e.g. Carey, De Schonen, & Ellis 1992).

To test whether this feature-by-feature approach enhanced unfamiliar face matching, Towler, White and Kemp (2017) asked participants to rate the similarity of each of 11 features (e.g. ears, jawline, eyes, mouth, nose) before deciding whether the faces were of the same or different people. Two experiments showed that this approach enhanced the performance of novice participants, and a third study found that forensic facial image examiners trained to use this technique were, compared to students, more accurate overall. Interestingly, and consistent with the results of a previous study (White et al., 2015), facial examiners were also found to be less impaired by image inversion. Inverted faces are particularly difficult to recognize, and this is thought to reflect the fact that recognition memory for faces is driven by holistic processing (e.g. Carey, De Schonen, & Ellis, 1992). The smaller inversion effect shown by experts is therefore suggestive that they are relying less on holistic and more on feature-based processing of the images. This is consistent with the training they receive, and also with the proposal that unfamiliar face matching tasks are driven by different perceptual processes than familiar face recognition (e.g. Megreya & Burton, 2006).

Super-recognizers: people with natural aptitude for face identification

In recent years, awareness of the difficulty of unfamiliar face matching tasks has extended outside of academia. With increased awareness of the problem, the onus has fallen on researchers to provide solutions that are robust enough to translate into tangible ‘real-world’ gains in accuracy, that can be used to improve the reliability of face identification in security and forensic professions.

One solution that has particular potential is to select individuals who are naturally good at face matching for these roles. The data from many of the studies reviewed in this chapter show striking variation in accuracy from one individual to the next, with some performing at close to chance levels while others are almost always correct. These wide ranges of accuracy have been highlighted in many studies, both in novices (Megreya & Burton, 2007) and professional populations (White et al., 2014, 2015; Wirth & Carbon, 2017). Whereas some individuals perform very poorly – barely

above chance, others perform extremely well – at 100% on standardized tests. Figure 5 in the previous section illustrates this point. While it is clear that professional experience does not predict accuracy, it is also clear that some passport officers performed very well in these tests and others performed very poorly.

People with extraordinary ability to recognize faces have been called 'Super-recognizers', based on the fact that their accuracy in face recognition tasks significantly exceeds that of typical individuals (Russel, Duchaine & Nakayama, 2009). Mounting evidence that individual differences in face identification are stable over time (Megreya & Burton, 2007) coupled with evidence showing that these abilities are hereditary (Wilmer et al., 2010; Shakeshaft & Plomin, 2015), has led to the proposal that selecting super-recognizers for professional roles can improve the accuracy in real-world face identification processes (White et al. 2014, Bobak, Dowsett & Bate, 2016; Noyes, Phillips & O'Toole, 2017). Indeed, large organisations are currently changing their recruitment policies in light of this discovery. For example, the Australian Passport Office (White, Dunn, Schmid & Kemp, 2015) and Metropolitan Police in London (Robertson et al. 2016; Davis, Lander, Evans & Jansari, 2016) have both established groups of individuals with superior accuracy in face identification tasks by using standardized tests of face identification ability developed by the scientific community.

At present, it is not clear how these 'super-recognizers' perform relative to the high performing forensic experts tested by White et al. (2015) and others. Studies do suggest that the type of cognitive processing producing superior performance in super-recognizers is qualitatively different from these forensic examiners. Specifically, feature-based analysis performed by forensic experts (see White et al., 2015) contrasts with the more holistic processing that appears to underpin super-recognizers' ability (see Russel et al., 2009; Bobak, Bennetts, Parris, Jansari & Bate, 2016). However, at the time of writing, a direct comparison of the accuracy of these groups has not been performed.

Given the impressive abilities of super-recognizers in the emerging literature on this topic (for a review see Noyes, Phillips, & O'Toole, 2017), it is important to ask whether these individuals should be allowed to provide expert testimony in court. Currently, courts rely on proof of training and experience when accrediting forensic image analysts as exert witnesses (Edmond et al. 2009), but do not require performance data showing that these accredited witnesses have superior face identification abilities. Instead of defining expertise in terms of experience and training – which appear to have limited effect on face matching accuracy – perhaps courts should instead demand proven accuracy on these tasks? This could be achieved, for example, by requiring face identification specialists to complete a standardized, empirically validated proficiency test in order to qualify as an expert witness. If this were to become the sole basis of claims to expertise, then super-recognizers would presumably qualify as expert witnesses.

This emerging literature has led some experts in evidence law to make the radical suggestion that groups of super-recognizers, established independently of police services, could in fact replace current face identification experts (Edmond & Wortley, 2016). This proposal requires careful consideration and will entail careful comparison of the relative merits of these types of testimony. For example, forensic facial

examiners provide detailed court reports comparing individual facial features and explaining the basis for their identification judgment. These analytic methods are compatible with cross-examination because they can be verbalized. At the moment it is not clear how super-recognizer testimony would achieve a similar level of transparency, as the basis of their superior abilities appears to rely on a more holistic and intuitive process. On the other hand, one could argue that their proven ability in the task is more important than the transparency afforded by standardized analytic processes.

Generating images of suspects

In the final section of this chapter we turn our attention to existing and developing approaches that allow investigators to construct a likeness of a person. This technology is used in cases where investigators do not have a suspect.

Reconstructing faces from memory

One of the first questions that police might ask eyewitnesses is, "can you describe the person you saw?". Simple verbal descriptions of an offender may be useful, but sometimes police will ask the witness to produce an image of the culprit. Historically, these likenesses were produced by a witness working alongside a police artist who would draw the culprit's face as the witness described it, but this has mostly been superseded by the introduction of facial composite systems.

Initially, composite systems consisted of a catalogue of drawings or photographs of possible features. For example the *Photofit* system developed by Jacques Penry included multiple photographs of each facial feature. The witness would sort thought the options for each feature to select the closest match to their memory and arrange the selected features on a face outline. Subsequently, these mechanical systems were replaced by computerized systems but these retained the same feature based approach. Early research on the quality of the likenesses produced was damning. Ellis, Davies, and Shepherd (1978) found that even when participants were able to study a photograph of the target while creating the composite, the resulting composite image bore little resemblance to the target. Further, a study by Christie and Ellis (1981) showed that drawings made by untrained mock witnesses were identified just as accurately as the composites they were able to produce.

This led to the development of a new generation of systems which were designed to build on psychological knowledge of face perception. As we discussed in the previous section, recognition memory for faces is known to operate by a holistic process that matches encountered faces to gestalt memory representations. As a result, the most recent composite systems have moved away from the feature-based *Photo-fit* approach. In "evolutionary" composite systems, such as EFIT-V and EvoFIT, witness descriptions are used to generate a small number of possible likenesses. The witness selects the best of these likenesses which is then used to generate, or "breed", a new set of likenesses. These systems enable witnesses to search a large mathematically defined space of possible faces by manipulating facial appearance holistically and refining the overall face template until it approximates their memory (see Frowd, Hancock, & Carson, 2004; Hancock, 2000; Solomon, Gibson, & Maylin, 2009).

There is some evidence that these developments have resulted in systems which can produce more accurate likenesses. A meta-analysis of 23 published studies of the likenesses produced using feature based and evolutionary systems found that evolutionary systems produced faces which were over four times more likely to be identified (Frowd et al., 2015). In a body of work spanning a decade, participants correctly named the individual pictured in the Evo-fit generated likeness on around 50% of cases. Indeed, one study (Frowd et al., 2013), found that when combined with an enhanced interview and a number of other techniques, the EvoFit system helped participant-witnesses produce likenesses that were correctly identified in 74% of cases. However, this study and the majority of other evaluations employed a relatively short retention interval of just 24 hours. Unsurprisingly, shorter retention intervals are associated with better quality composites (Frowd et al., 2015), and so these lab-based estimates may overestimate operational accuracy.

Is there anything we can do to further enhance the quality of the likenesses produced by witnesses? One interesting possibility is to 'fuse' multiple likenesses. In some cases a culprit may be seen by several independent witnesses, either in the commission of a single offence or multiple offences. In these cases the police may end up with several different likenesses. How can we best use these multiple images? Brace et al. (2006) investigated whether Police should publish more than one of these images, or whether differences in appearance would confuse viewers. In their study, two groups of witnesses watched a mock crime and then worked with trained police composite operators to produce a likeness, resulting in eight likenesses of each of two culprits. These images were then shown in sets of 1, 4 or 8 composites to participants who were familiar with the culprits. Showing more than one image was found to increase identification rates. Interestingly, if only one composite could be published, then the authors found that the one which looked, on average, most like the others in the set was the image that would give rise to the best identification rates.

An alternative, but conceptually similar approach, is to combine composites produced by different witnesses by digital morphing. Using this approach, Bruce et al. (2002) found that the resulting average was rated as a better likeness than single composites, on average, and as good as the best composite. A similar pattern of results emerged when participants used the composite to try to select the culprit's photograph from an array. Taken together, the results of these two studies suggest that it may be possible to leverage some additional value from the composites by aggregating likenesses produced by independent witnesses. These findings are also in line with studies of unfamiliar face matching showing that aggregating identification judgments made by independent viewers (White et al., 2013, 2015) and presenting multiple images of the target (White et al., 2014) enhances matching accuracy.

Composites without witnesses

Perhaps in the future it may even be possible to construct facial composites without the involvement of witnesses. Facial appearance is largely determined by our DNA. We see evidence of this every time we look into the faces of members of our family; we look like our close relatives, with identical twins providing the clearest demonstration of this fact. As a result, geneticists are currently working on methods to

construct a likeness of a person from a sample of their DNA, raising the possibility that this could be used to generate images of suspects in police investigation.

Until recently the idea of "genetic photofitting" -- or more formally "Forensic DNA Phenotyping" -- sounded like science fiction, but it is now a rapidly advancing field of research. For some time geneticists have been able to identify genes controlling certain basic facial characteristics, such as eye and hair color (Kayser, 2015). Callaway (2009) describes how forensic scientists investigating the 2004 Madrid train bombing analyzed DNA samples recovered from a toothbrush found at an apartment used by the bombers. The DNA recovered from the toothbrush did not match any known suspects, but analysts were able to determine that the sample was likely to belong to a person originating from North Africa rather than Europe, a finding which helped investigators identify the likely terrorists.

More recently, researchers have begun to identify genes which control the structural appearance of the face. Liu et al. (2012) examined three dimensional shape data from MRI scans and photographs of almost 10,000 people of European origin. Using this data they selected 9 facial "landmarks", including the location of the left and right eyes and the bridge and tip of the nose. The authors were then able to identify five genes which were associated with the location of these landmarks in a predictable way. For example, variation in gene PRDM16 was associated with changes to the nose width and length, while variants of gene TP63 were associated with changes in the distance between the eyes. This work was advanced further by a team of researchers who collected DNA and 3-D face scans of 592 individuals of European and African ancestry and identified 20 genes associated with changes in face shape (Claes et al. 2014a; 2014b). Exploring the forensic implications of this work, Claes et al. (2014b) used this database to generate facial composites from DNA. This work remains exploratory, and the resemblance between likenesses generated by DNA modelling and the individuals' actual appearance were not compelling. It appears that much of the variation in appearance, for this population at least, could be predicted from ancestry and sex information alone.

Nevertheless, genetics research is expanding at a rapid rate, and progress in this area is accelerating in tandem with available computing power. This may lead to "genetic photofitting" becoming a viable technology in the future, with investigators estimating appearance of a suspect based on a DNA sample found at the crime scene. This leads to some important psychological and legal questions. How will the investigators use this information? For example, will they publish these images in the hope that someone will note a similarity to someone they know, or will it be possible to use the composite as a template to search databases of images sourced from government identity documents or CCTV?

Imagine such a search returns your face. Who will decide whether you match the genetic photofit, and how will they make this determination? Will this similarity in appearance be sufficient evidence to require you to provide a DNA sample for testing? Another intriguing possibility is that the genetic photofit may be combined with other sources of information. For example it may be possible to fuse the DNA and eyewitness-derived likeness – either by averaging these images (Bruce et al. 2002) or perhaps by using the DNA photofit as a prompt for the eyewitness or as a starting point for the latest generation of composite systems such as Evo-fit. These

suggestions will of course require careful empirical testing, but for now they are a further illustration of how current and future technologies change, but do not eliminate, human involvement in the process of identifying unfamiliar faces.

CONCLUSIONS AND FUTURE DIRECTIONS

In this chapter we have outlined how face identification from images impacts multiple stages of the legal system. This task plays an important role in protecting against identity fraud and in identifying people in day to day life, but the accuracy of photo identification is less reliable than one would hope or expect. This limitation of human perception causes vulnerabilities at many stages in the legal system: is this person who they claim to be? Does the suspect match the culprit on CCTV? Is the person depicted on this police wanted poster someone I know? These are all critical decisions in legal processes that rely on people's ability to identify faces from images. As we have seen, errors in these tasks are prevalent in both novices and experts.

The nature of human involvement in face identification systems is changing at a rapid rate. Deployment of face recognition technology, combined with increased use of digital imagery in casework, is producing qualitative changes in the operation of legal identity verification. All too often, policy plays catch-up with technological change and face recognition is no exception. Evidence suggests that these new tools in fighting crime may also pose an increased risk of false identification evidence being presented in court. In this context, it is critical that modern systems are designed in a way that balances computer and human processing to optimize the accuracy of these systems (Towler, Kemp, & White, 2017), and to ensure that they are not biased towards incriminating minority groups. Human decision-making is a critical component of this new identification paradigm and successful implementation of these systems demands a thorough understanding of human performance on these tasks.

Responsibility therefore falls on psychologists and vision scientists to develop a theoretical understanding of the mechanisms driving high levels of performance. This can help, for example, to ensure that face recognition software is adjudicated by human users who are selected on the basis of their ability to perform the task accurately. In light of evidence showing stable individual differences in novice and expert populations, and genetic studies showing that this ability is largely hereditary, it appears likely that this solution can help minimize false identifications. However, current definitions that are used by courts to decide whether to admit expert testimony will need to be revised before super-recognizer testimony can be utilized effectively (Edmond &Wortley, 2016). Proper treatment of these legal issues will require a better theoretical understanding of superior face identification abilities than what we currently have. Super-recognizers may yet offer the *deus ex machina* to many problems raised in this chapter, but scientific knowledge on this topic is in its infancy.

Finally, in developing our understanding of expertise in identifying people from images, it may be necessary to consider the question more broadly than we have in this chapter. Despite the title of this chapter being concerned with identifying *people*, we have focused exclusively on the task of identifying *faces*. We have adopted this focus due to space constraints and also because the vast majority of research in this area is focused on face identification. Nevertheless, when identifying people from

video, identification often stems from, or is supported by, sources of information beyond the face: a person's body shape, gait and clothing all provide potentially useful cues to identity (Hahn, O'Toole, & Phillips, 2016; Yovel & O'Toole, 2016). Currently it is not clear whether expertise in face identification also entails superior abilities in processing these other channels of information. Future work that aims to develop our understanding of expertise in person identification more broadly can help to improve the accuracy, interpretation and transparency of identification evidence in the legal system.

AUTHOR NOTE

The answer to the face matching array in Figure 1 is target absent.

References

Attwood, A. S., Penton-Voak, I. S., Burton, A. M., & Munafò, M. R. (2013). Acute anxiety impairs accuracy in identifying photographed faces. *Psychological Science*, *2*4, 591-1594.

Beattie, L., Walsh, D., McLaren, J., Biello, S. M., & White, D. (2016). Perceptual impairment in face identification with poor sleep. *Royal Society Open Science*, *3*, 160321.

Bertillon, A. (1893). *Identification anthropométrique: instructions signalétiques* (Vol. 1). Impr. administrative.

Bobak, A. K., Bennetts, R. J., Parris, B. A., Jansari, A., & Bate, S. (2016). An in-depth cognitive examination of individuals with superior face recognition skills. *Cortex*, *82*, 48-62.

Bobak, A. K., Dowsett, A. J., & Bate, S. (2016). Solving the border control problem: evidence of enhanced face matching in individuals with extraordinary face recognition skills. *PloS One*, *11*, e0148148.

Brace, N., Pike, G., Kemp, R., Turner, J., & Bennett, P. (2006). Does the presentation of multiple facial composites improve suspect identification? *Applied Cognitive Ppsychology*, *20*, 213-226.

Breen, N. (2009, March 22). Pauline Hanson: We're sorry, the nude photos weren't you. *The Daily Telegraph.* Retrieved from http://www.dailytelegraph.com.au/news/national/pauline-were-sorry-they-werent-you/news-story/ff13b1b8d6a519d80dba03c42c7dff99

British Security Industry Associations (2015). *The picture is not clear: How many CCTV surveillance cameras in the UK?* Industry report. Accessed from: https://www.bsia.co.uk/publications/publications-search-results/195-the-picture-is-not-clear-how-many-cctv-surveillance-cameras-in-the-uk.aspx (8 August, 2017).

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., Ness, H., Hancock, P. J., Newman, C., & Rarity, J. (2002). Four heads are better than one: combining face composites yields improvements in face likeness. *Journal of Applied Psychology*, *87*, 894.

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., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*, 286-291.

Burton, A. M., Schweinberger, S. R., Jenkins, R., & Kaufmann, J. M. (2015). Arguments against a configural processing account of familiar face recognition. *Perspectives on Psychological Science*, *10*, 482-496.

Calloway, E. (2009). DNA mugshots' narrow search for Madrid bombers. *New Scientist*.

Campbell-Tiech, A. (2005). "Stockwell" revisited: The unhappy state of facial mapping. *Archbold News*, *6*, 4-6.

Carey, S., De Schonen, S., & Ellis, H. D. (1992). Becoming a face expert. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *335*, 95-103.

Christie, D. F., & Ellis, H. D. (1981). Photofit constructions versus verbal descriptions of faces. *Journal of Applied Psychology*, *66*, 358.

Claes, P., Liberton, D. K., Daniels, K., Rosana, K. M., Quillen, E. E., Pearson, L. N., ... & Tang, H. (2014a). Modeling 3D facial shape from DNA. *PLoS genetics*, *10*, e1004224.

Claes, P., Hill, H., & Shriver, M. D. (2014b). Toward DNA-based facial composites: preliminary results and validation. *Forensic Science International: Genetics*, *13*, 208-216.

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating Predictors of Superior Face Recognition Ability in Police Super-recognisers. *Applied Cognitive Psychology*, *30*, 827-840.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*, 482-505.

DesLauriers, Richard (2013). "Remarks of Special Agent in Charge Richard DesLauriers at Press Conference on Bombing Investigation" (press release). Boston: FBI. Retrieved August 8, 2017 from: https://archives.fbi.gov/archives/boston/press-releases/2013/remarks-of-special-agent-in-charge-richard-deslauriers-at-press-conference-on-bombing-investigation-1

Edmond, G., Biber, K., Kemp, R. I., & Porter, G. (2009). Law's looking glass: expert identification evidence derived from photographic and video images. *Current Issues in Criminal Justice*, *20*, 337.

Edmond, G., Kemp, R., Porter, G., Hamer, D., Burton, M., Biber, K., & Roque, M. S. (2010). Atkins v The Emperor: the 'cautious' use of unreliable 'expert'opinion. *The International Journal of Evidence & Proof*, *14*, 146-166.

Edmond, G., & Wortley, N. (2016). Interpreting Image Evidence: Facial Mapping, Police Familiars and Super-Recognisers in England and Australia. *Journal of International and Comparative Law*, 3, 473-522.

Ellis, H. D., Davies, G. M., & Shepherd, J. W. (1978). A Critical Examination of the Photofit System For Recalling Faces. *Ergonomics*, *21*, 297-307.

Facial Identification Scientific Working Group (2012). *Guidelines for Facial Comparison Methods*. Retrieved from: https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf (1 August 2018)

Frowd, C. D., Erickson, W. B., Lampinen, J. M., Skelton, F. C., McIntyre, A. H., & Hancock, P. J. (2015). A decade of evolving composites: regression-and meta-analysis. *Journal of Forensic Practice*, *17*, 319-334.

Frowd, C. D., Hancock, P. J., & Carson, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Perception (TAP)*, *1*, 19-39.

Frowd, C. D., Skelton, F., Hepton, G., Holden, L., Minahil, S., Pitchford, M., ... & Hancock, P. J. (2013). Whole-face procedures for recovering facial images from memory. *Science & Justice*, *53*, 89-97.

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society Open Science*, *4*, 170249.

Garvie, C., Bedoya, A. M., & Frankle, J. (2016). *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law Center on Privacy & Technology. Retrieved from: www.perpetuallineup.org (1 August, 2017)

Grother, P., & Ngan, M. (2014). Face Recognition Vendor Test (FRVT). *Performance of Face Identification Algorithms, NIST Interagency Report*, *8009*, 84.

Hahn, C. A., O'Toole, A.J. & Phillips, P. J. (2016). Dissecting the time course of person recognition in natural viewing environments. *British Journal of Psychology, 107, 117-135*.

Hancock, P. J. (2000). Evolving faces from principal components. *Behavior Research Methods*, *32*, 327-333.

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

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*, 445-464.

Jenkins, R., & Burton, A. M. (2011). Stable face representations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *366*, 1671-1683.

Jenkins R., White D., van Montfort X., Burton A. M. (2011). Variability in photos of the same face. *Cognition, 121*, 313–323.

Kayser, M. (2015). Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes. *Forensic Science International: Genetics*, *18*, 33-48.

Kemp, R. I., 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 sciences*, *52*, 779-783.

Klontz, J. C., & Jain, A. K. (2013). A case study on unconstrained facial recognition using the boston marathon bombings suspects. *Michigan State University Technical Report 119*.

Lee, W. J., Wilkinson, C., Memon, A., & Houston, K. (2009). Matching unfamiliar faces from poor quality closed-circuit television (CCTV) footage. *Axis: The Online Journal of CAHId*, *1*, 19-28.

Leys, N. (2009, March 22). Majority of experts say photos not Pauline Hanson. *The Australian.* Retrieved from http://www.theaustralian.com.au/news/latest-news/photos-arent-of-hanson-paper-says/news-story/368cdffbe30296594df7e4d0a322cefa

Liu, F., Van Der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., ... & Van Der Lugt, A. (2012). A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS genetics*, *8*, e1002932.

Megreya, A, & 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 & Psychophysics*, *69*, 1175-1184.

Megreya, A., & Burton, A. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*, 364–372.

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*, 700-706.

Megreya, A., White, D., & Burton, A. M. (2011). The other race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology, 64*, 1473-1483.

Meissner, C.A., & Brigham, J.C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology Public Policy and Law 7(1):*3-35.

Montgomery, D., Horwitz, S. and Fisher, M (2013, April 20) Police, citizens and technology factor into Boston bombing probe. *The Washington Post*. Retreived from: www.washingtonpost.com/world/national-security/inside-the-investigation-of-the-boston-marathon-bombing/2013/04/20/19d8c322-a8ff-11e2-b029-8fb7e977ef71_story.html?utm_term=.c5712a87dacc

Norell, K., Läthén, K. B., Bergström, P., Rice, A., Natu, V., & O'Toole, A. (2015). The effect of image quality and forensic expertise in facial image comparisons. *Journal of Forensic Sciences*, *60*, 331-340.

Noyes, E., & Jenkins, R. (2016). Deliberate disguise in facial image comparison. *Journal of Vision*, *16*, 924-924.

Noyes, E., Phillips, P. J., & O'Toole, A. J. (2017). What is a super-recognizer? In *Face Processing: Systems, Disorders and Cultural Difference*,. Bindemann, M. & Megreya, A.M. (Eds.), Nova Science.

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition*, *165*, 97-104.

Yovel, G., & O'Toole, A. J. (2016). Recognizing People in Motion, *Trends in Cognitive Sciences*, *20*, 383-395.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). Another-race effect for face recognition algorithms. *ACM Transactions on Applied Perception*, *8*, 14.

Phillips, P. J., Hill, M. Q., Swindle, J. A., & O'Toole, A. J. (2015). Human and algorithm performance on the PaSC face recognition challenge. In: *IEEE 7th International Conference on Biometrics: Theory, Applications and Systems (BTAS 2015).*

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, *141*, 161-169.

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police super-recognisers. *PloS one*, *11*, e0150036.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic bulletin & review*, *16*, 252-257.

Shakeshaft, N. G., & Plomin, R. (2015). Genetic specificity of face recognition. *Proceedings of the National Academy of Sciences*, *112*, 12887-12892.

Solomon, C., Gibson, S., & Maylin, M. (2009). A new computational methodology for the construction of forensic, facial composites. *Computational forensics*, 67-77.

Spiteri, V. R., Porter, G., & Kemp, R. (2015). Variation of craniofacial representation in passport photographs, *Journal of Criminological Research, Policy and Practice*, 1, 239-250.

Strathie, A., McNeill, A., & White, D. (2012). In the dock: Chimeric image composites reduce identification accuracy. *Applied Cognitive Psychology*, *26*, 140-148.

Strathie, A., & McNeill, A. (2016). Facial Wipes don't Wash: Facial Image Comparison by Video Superimposition Reduces the Accuracy of Face Matching Decisions. *Applied Cognitive Psychology*, *30*, 504-513.

Towler, A. (2017) Match me if you can: Evaluating professional training for facial image comparison (unpublished PhD thesis). UNSW Sydney, Sydney, Australia

Towler, A., Kemp, R. I., & White, D. (2017). Unfamiliar face matching systems in applied settings. In M. Bindemann & A. M. Megreya (Eds.), *Face Processing: Systems, Disorders and Cultural Difference,*. Nova Science.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*, 214-218.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*, 47-58.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology*, *27*, 769-777.

White, D., Burton, A. M., Jenkins, R. & Kemp, R. I. (2014). Redesigning photo-ID to improve unfamiliar face matching. *Journal of Experimental Psychology: Applied*, 20, 166-173.

White, D., Dunn, J. D., Schmid, A. C. & Kemp, R. I. (2015a). Error rates in users of automatic face recognition software. *Plos One* 10: e0139827.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS One*, *9*, e103510.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A.J. (2015b). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society of London B: Biological Sciences*, 282, 1814-1822.

White, D., Norrell, K., Phillips, J. P., O'Toole, A. J. (2017). Human factors in forensic face identification. In *Springer Handbook of Biometrics in Forensic Science*, Tistarelli, M. & Champod, C. (Eds.), Springer-Verlag.

Wilkinson, C., & Evans, R. (2009). Are facial image analysis experts any better than the general public at identifying individuals from CCTV images? *Science & Justice*, *49*, 191-196.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., ... & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of ciences*, *107*, 5238-5241.

Wirth, B. E., & Carbon, C. C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied*, *23*, 138.

Woodhead, A. D. Baddeley & D. C. V. Simmonds (1979). On Training People to Recognize Faces, *Ergonomics*, *22*, 333-343.