Automatic recognition systems and human computer interaction in face matching.

Eilidh Noyes[1*] & Matthew Q. Hill[2]

[1] Department of Psychology, University of Huddersfield, United Kingdom.
[2] School of Behavioural and Brain Sciences, The University of Texas at Dallas, Richardson, TX, US.

Abstract

The human face facilitates identification in security and policing scenarios. Automatic face recognition systems have increased in prevalence and accuracy in recent years. As a result, the identification task, which once fell entirely to humans, is now a process performed by man and machine. Automatic face recognition systems provide image similarity comparisons and can create candidate lists to narrow down potential targets. The design, operational usage and effectiveness of these automatic systems, as well as the interaction of human and computer recognition are the topics of this chapter.

Keywords: face recognition, automatic face recognition, automatic face recognition, deep convolutional neural networks, human computer interaction.

*Corresponding Author address: Dr. Eilidh Noyes, Department of Psychology, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, UK. Email: e.noyes@hud.ac.uk

Introduction

Automated face recognition systems play an important role in security and policing scenarios. They can process large data at rapid speeds, and, unlike humans, are not affected by limiting factors such as fatigue or boredom (Alenezi, Bindemann, Fysh, & Johnston, 2015; Beattie, Walsh, McLaren, Biello, & White, 2016). Their prevalence has increased in recent years, which is a direct reflection on advances in both algorithm technology and consequently in performance.

At the present time, the role of face recognition algorithms is typically to *assist* humans in the identity verification process. Whilst an algorithm can unlock personal electronic devices with a high degree of autonomy, in security and policing scenarios, identification systems generally involve both humans and machines.

The example that is likely most familiar to readers is the use of electronic e-gates at passport control. E-gates make identifications by comparing a traveller's live image against their passport image (1:1 image matching) for similarity. If a potential mismatch in identity is detected, the traveller is referred to a human operator who will adjudicate the identification.

Human and machine face recognition systems are also used to assist identifications in policing. Computer algorithms aid suspect identification by comparing a target image for similarity against a database of known offenders. The images of highest similarity are returned as a 'candidate list', which is subsequently reviewed by a human officer. Algorithms can compare the target image against far more images than possible by human eye alone. The number of face images present in a comparison database can range from just a few, to millions. Algorithms can process image similarity computations in fractions of a second.

Where many images of multiple people are of interest, algorithms can facilitate grouping of identities (clustering). Again, a human review is required to verify algorithm accuracy.

This chapter is divided into four sections. In Section 1 we provide a brief overview of how algorithms work. In Section 2 we discuss the role that algorithms play in Policing and Border Control scenarios. In Section 3 we review accuracy of automatic face recognition systems and provide comparisons with human performance. Finally, in Section 4 we consider the advantages and disadvantages of human and machine interaction.

Section 1. A basic overview of the workings of algorithms

The design and structure of automatic face recognition systems is fascinatingly complex. However, basic knowledge of algorithm design and the image comparison procedure allows the strengths, limitations, and potential of these systems to be evaluated. The following text provides a comprehensive overview of algorithm design and the face identification procedure.

How do algorithms determine identity from images?

Algorithms perform a series of steps to compute image comparisons. First, they must find the face in the image (face detection). Next, the algorithm must process relevant features of the image (feature extraction), and configure a measure of similarity between images (distance metric learning). A similarity score can then be computed and compared against a decision rule (used to determine whether the images are similar enough to be considered the same identity). There are different ways of extracting features from an image. These will be looked at in more detail below, however first we will consider the comparison metric used by most algorithms, which is an image *similarity* score. A high similarity score for a pair of images indicates that they are likely to be images of the same identity, whereas a low similarity score suggests that the images are of different identities.

This similarity score can be used to aid identification in 1:1 comparison, 1:N comparison, and clustering scenarios.

- 1:1 image matching: the similarity score helps determine if two specific images are of the same identity or different identities. For example, the passport control e-gate scenario whereby a passenger's live-capture image is compared against the passport photo.
- 1:N image matching: similarity scores are obtained for a target image against a database of other images. Typically, a specified number of images of highest similarity to the target are returned as a ranked list. For example, the image of a suspect (captured by CCTV) can be compared against a database of police custody images of known offenders.
- Clustering: a large set of images can be grouped according to similarity scores. A set of rules are used to determine which images in a set are likely to be of the same person based on their similarity to other images in the set.

But how similar must the images be to be considered the same identity? An operator defined criterion score determines the level of similarity required to constitute an identity match. An algorithm can make two types of error— false acceptance or false rejection. The criterion score dictates the likelihood of each error type. Choosing a criterion for an imperfect system requires a trade-off between the two types of error (false alarm, false reject).

- **False acceptance rate**: this is how often an algorithm mistakes two different people as being the same person. The algorithm falsely interprets the images as being the same identity. E.g. accepting a stolen passport as genuine.
- **False reject rate:** this is how often the algorithm fails to match images of the same face. E.g. not accepting a genuine passport holder as a match to their passport image.

This criterion score is based on operational false acceptance rates— the score at which *x* percentage of non-matched identity pairs will be labelled as a matched identity pair by the network. In practice, the criterion score for a system often depends on its operational usage. A shift in criterion changes the likelihood of a false match scenario (saying that two people are the same when they are actually different identities, e.g., accepting an imposter as a match to a stolen passport) versus that of a false reject (e.g., saying that a person is not a match to their passport when they

actually are) . The trade-off must be considered carefully for operational usage (see Figure 1). If there is a greater risk associated with accepting two different people as the same identity than there is of rejecting multiple images of a true same identity as a match, then a more conservative criterion should be chosen (lower false accept rate, higher false reject rate). On the other hand, if it is more risky to reject multiple images of the same person, then a more liberal criterion score should be used (higher false accept rate, lower false reject rate).

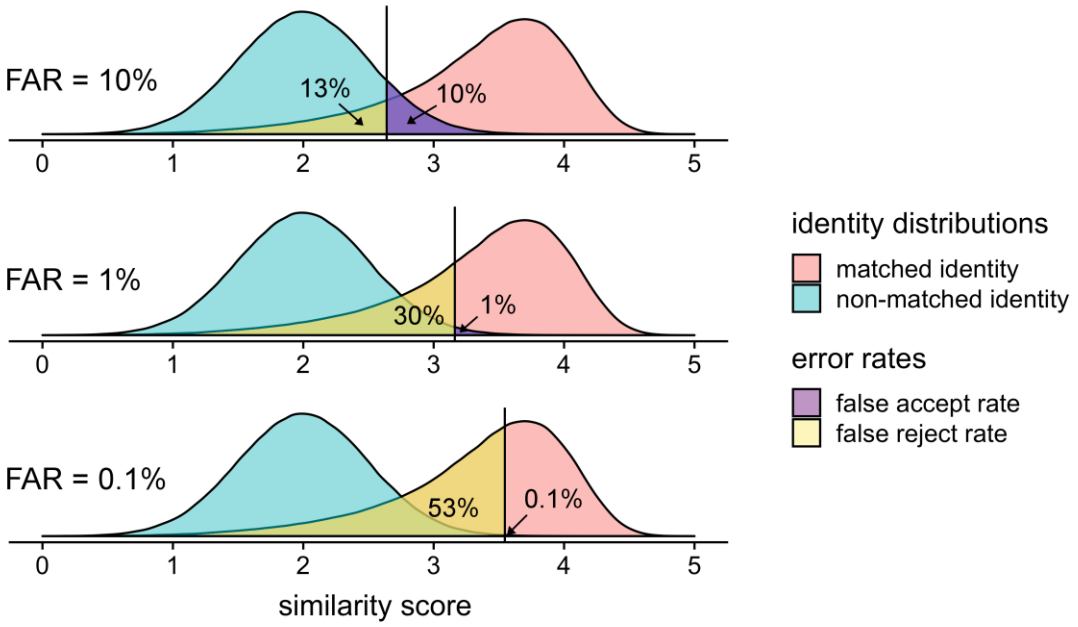

Figure 1. Choosing a criterion for an imperfect system requires a trade-off between the two types of error (false alarm, false reject). In this example, setting the criterion to a 10% false accept rate (FAR) results in a false reject rate (FRR) of 13%. A more conservative FAR of 1% results in an FRR of 30%. The FAR of 0.1% that is required by passport systems results here in an FRR of 53%. This is significantly higher than the allowed 5% FRR of passport systems, meaning this example system would not be accurate enough for their standards.

Algorithm Design

Below we provide an overview of early algorithm designs and more modern designs.

Early Algorithm Design

Early algorithms were programmed to compute image similarity scores or identity classifications using *hand-coded* features (e.g., lines in the image) on a pixel-by-pixel basis. In other words, the code dictates the image qualities that are compared by the automatic system. To illustrate this process with an oversimplified example, an algorithm might be programmed to compare the location of black pixels in one image with black pixels in a second image. For the purpose of this example, a strong match in black pixel location is indicative of an identity match. In practice, the feature extraction procedures are far more sophisticated.

Early face recognition algorithms generally consisted of two stages of processing: feature extraction, in which the algorithm processes information in the image deemed useful for the task; and distance metric learning, in which the algorithm tunes up a similarity measure to compare the extracted features of two images. The types of

features extracted from an image in the first stage were often carefully designed, and based on well-known principles of cognition or information theory. Many of these techniques employed a process known as convolution. Convolution uses a matrix of numbers called a 'kernel' that describes a pattern to be detected in an image (see Figure 2). This kernel is 'convolved' over the image, tiling in steps of *n* pixels, where *n* is a parameter chosen by the designer of the algorithm. The output is a matrix of either the same size as the original image, or smaller, depending on the step size *n*. The values of the output matrix can be treated as a measure of how well the pixels at a given location in the image match the pattern (i.e., feature) detailed by the kernel. This makes the output matrix itself equivalent to a map of the feature's presence in the image. The fact that convolution outputs not only the presence of a feature, but also a measure of its location in the image, makes it a powerful method of feature extraction. However, this strength can also be a weakness. This method's reliance on location means it requires carefully aligned stimuli in order to match patterns effectively.

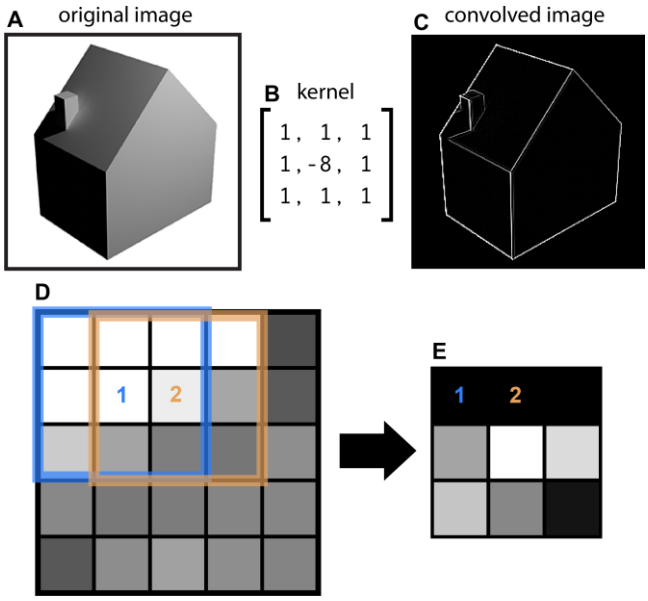

Figure 2. In the process of convolution, (A) an image is taken as input. (B) A matrix of numbers called a "kernel" defines the feature to be detected. Here, a common edge detection kernel is shown. (C) The final result is an image in which high values (white) represent a close match with the kernel's feature at that location, and low values (black) represent the absence of the feature. (D) To find the feature over the whole image, the kernel is tiled in steps of one or more pixels. The blue box (1) and yellow box (2) show the area covered by the first and second steps, respectively. (E) Each cell in the kernel is multiplied by the value at its overlapping pixel in the image at each location. The multiplied numbers are then summed to give the final result at that location. The blue 1 and yellow 2 show the result of the convolution from steps one and two respectively. These pixels are black because the kernel was not a good match with the image at those locations.

The second stage—distance metric learning—involves choosing or building a measure of similarity in which face images can be compared. An early distance metric used for computer face recognition was Principal Component Analysis (PCA) (Turk & Pentland, 1991), a technique also used for image compression. This method reduces the size of the face representation while preserving important information, which made it an attractive solution in a time when storage space and processing power were scarce. PCA is an 'unsupervised' learning method, meaning that it is

trained without any information about true identity labels. Some other methods use training data labelled with the true identity in order to improve performance (e.g., Linear Discriminant Analysis and Support Vector Machines). These methods are referred to as 'supervised'. Once a distance metric is in place, the operator can obtain similarity scores between sets of images.

Deep Convolutional Neural Networks

The current state-of-the-art in face recognition algorithms is defined by Deep Convolutional Neural Networks (DCNNs). Consistent with some of their predecessors, these networks use convolution operations to extract features from an image. The main innovations of DCNNs are twofold: 1.) they leverage multiple layers of processing (the *depth* in a *deep* neural network); and 2.) the (multiple) kernels they use at each layer are not hand-coded, but rather learned from training data (Krizhevsky, Sutskever, & Hinton, 2012). The depth of these networks allows them to combine simple features from early layers into more complex features at deeper layers. This is similar to how the primate brain processes visual information, and gives DCNNs the ability to process images robustly across significant changes in viewpoint and illumination (Cadieu et al., 2014).

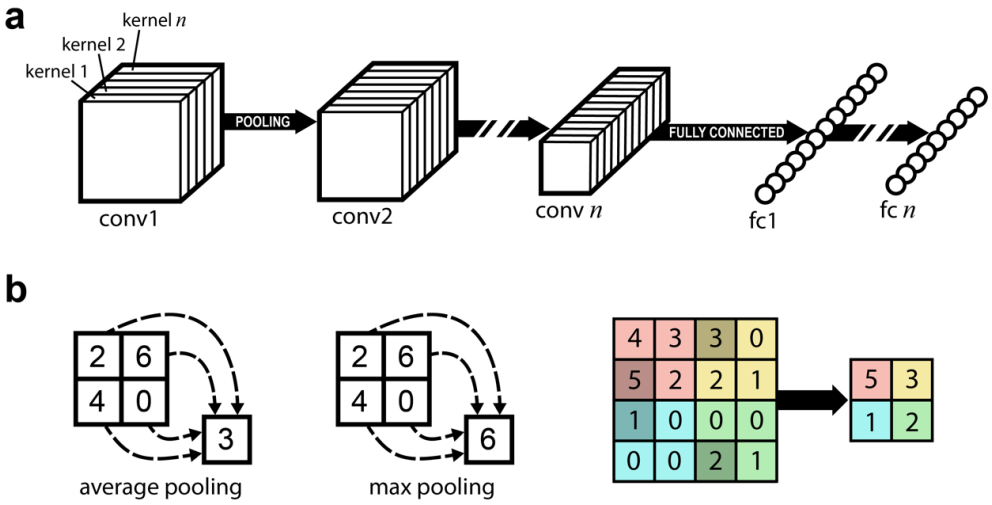

Figure 3. (A) The architecture of a DCNN can take many forms, but the common structure is such that multiple layers of convolution and pooling are followed by one or more "fully connected" layers. Each convolutional layer utilizes multiple kernels (see Figure 2) to find features in the image. The image is processed from left to right, where "fc *n*" produces the final identity representation. (B) Pooling combines the values of neighboring pixels in order to reduce the size of the representation. This also increases the size of a kernel's receptive window relative to the original image. Average pooling (left) takes a contiguous block of pixels and averages them together. More common in DCNNs is max pooling (middle), which simply chooses the highest value in the block as its output. This allows the strongest signal of a feature within that block to pass through the network. A simple example of max pooling (right) is shown for illustration. Each color shows a pooling block, with the highest value darkened for emphasis.

Each layer in a deep network can be thought of as one of many consecutive steps in processing (see Figure 3). The first layer of an effective network will capture simple features in the image such as lines, dots, or opposing colours (cf., Zeiler & Fergus, 2014). The second layer's features are created by combining the first layer's features

into more complex features. These second-order features might, for example, combine lines at different orientations into corners or curves. Each layer builds its features by combining the features of the previous layer. After one or more layers of convolution, a pooling layer will combine neighbouring pixels in groups of four (i.e., 2×2) or more, depending on the architecture of the network. Pooling reduces the size of a layer while also increasing the receptive window of a feature. This means that a feature in a deeper layer of a DCNN will represent information from a larger section of the original image than an earlier-layer feature. As a consequence of this, the deeper layers of a DCNN are less bound to specific locations in the original image than early layers. After multiple layers of convolution and pooling, a DCNN has one or more 'fully connected' layers. As the name suggests, each unit of this type of layer is fully connected to each unit of the previous layer. Here, a unit refers to a simulated neuron, or a node connected to other nodes in the neural network. Unlike convolutional layers, the units of a fully connected layer have no relationship to a location in the original image.

When a face image is input into a face identification DCNN, the output is generally a vector of numbers that acts as an identity descriptor. This is conceptually similar to the features derived by traditional approaches, but derived entirely from the training images and identity labels. The angle (or sometimes distance) between two such vectors can be used as a measure of similarity between images. In order for this similarity score to result in a match or non-match decision about the two images, the designer must choose a criterion cut-off score. As described earlier, the criterion is the similarity score above which the images are considered a match, and below which the images are considered a non-match. One method for finding a criterion score is to choose the percentage of false positives an operator is willing to accept, and to find the similarity score that yields the desired percentage using a dataset with ground-truth identity labels.

The complexity of DCNNs makes them incredibly powerful, but it also requires a staggering amount of training data to be effective. For instance, the MegaFace training dataset contains 4.7 million images of 672 thousand different identities (Nech & Kemelmacher-Shlizerman, 2017). It is important that a training dataset not only have a large number of images, but also have multiple images per identity. The images should also be highly variable in terms of lighting, viewpoint, etc., so the network can properly learn to recognize a person across these changes. As a consequence of the complexity of the networks and the size of the training data, DCNNs can sometimes take weeks to train even with powerful parallel processing. Though once the network is trained, a face can be processed in fractions of a second.

Section 2. What role can algorithms play in security scenarios?

Automated Border Control and Secure Authentication (1:1 matching)

One of the most familiar uses of face recognition for people that travel internationally are Automated Border Control (ABC) e-gates. The traveller presents their passport to the e-gates, and faces the camera, allowing an image matching comparison to be computed. E-gates are featured at airports internationally. Best practice guidelines state that 'all ABC systems must be monitored by a human operator' (FRONTEX, 2015).

E-gates generally compute identification quickly, despite several steps involved in the identification process. The ABC system must first check that the passport chip is genuine and that it is a match to demographic information on the passport. The live capture image of the holder (taken by the camera attached to the e-gate) is then compared against the image stored on the passport chip. It is recommended that the captured image should be compared against the scanned image on the passport (to catch instances of passport tampering). Additionally, in an attempt to identify wanted suspects or missing people, these images may be compared against a database (e.g., a watch list). If there are any consistencies in image matches, a flag against a watch list image, or if the system 'times out', the traveller is referred to a passport officer for adjudication. Otherwise, the e-gate opens to allow the passenger to pass.

According to the FRONTEX guidelines, algorithms must operate at a false acceptance rate (FAR) of 0.1% or less (see Figure 4). This is equivalent to a scenario whereby someone gets away with using another person's passport (the e-gate opens in error) 1 in 1000 times. Additionally, the false rejection rate (FFR) should not exceed 5%. This means that only 1 in 20 people will be flagged in error to be processed by a border official. Interestingly, these recognition rates are quite liberal in comparison to those recommended for iris recognition rates (FAR = 0.001%). Independent tests of the ABC system are advised, and should preferably take place in the live environment as these may present new challenges which are not always included in algorithm testing procedures.

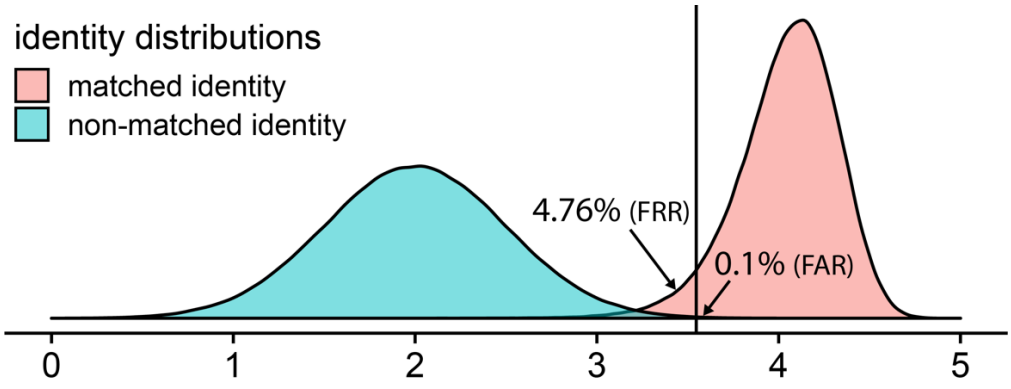

Figure 4. Hypothetical match and non-match distributions of a recognition system that would barely pass the criteria of the FRONTEX guidelines for passport recognition systems (FAR <= 0.1%, FRR <= 5%).

It is recommended that ABC e-gate systems should include 'live-ness' detection systems to check that a person has actually presented their face to the e-gate camera rather than a photograph or a mask (see Sanders & Jenkins, this volume) Camera quality is recommended to be at least 2 megapixels, with a minimum frame rate of 10 frames per second. Suggested lighting conditions are also reported in the FRONTEX report. As the above are recommended, rather than required, it follows that image capture systems may differ substantially across location.

Policing Scenarios (1:N matching)

In policing scenarios, an officer may be tasked with trying to identify a person from an image, such as a CCTV image from a crime scene. This can be a difficult task, and

there may be a large database of known offenders to compare the image against. An algorithm can *assist* the investigation by narrowing down the number of possible identity matches in a database. The target image is compared against database images, and the most similar images returned in a candidate list—a list of the most similar *x* number (number determined by the human operator) of identities to the input image (see Davies, Innes, & Dawson, 2018). Whilst it is time consuming for an officer to search a large database of potential suspects, an algorithm can perform such a task very quickly. As algorithms are not 100% accurate, a human operator is required to view the output image(s) and decide if a match is present.

Several police forces have trialed face recognition technology. However, there are few reports of the effectiveness of these systems in practice. Often, the ground truth for face identifications is unknown, making it difficult to assess the accuracy of these systems in practice. In the few reports that exist, the human operator's verdict is often taken as ground truth. This is problematic, as we know that human face matching is error prone.

Davies et al. (2018) provide a report of South Wales Police Usage of Automated Face Recognition Systems. Their report explains two uses of algorithms in police investigations. These are termed 'Locate' and 'Identify'.

The Locate mode is used in live face matching situations. A camera scans a live crowd and compares the faces against custody images from a police database. If a match is identified, the police officer is alerted. If the officer believes that the match is 'true', then the person may be stopped by an officer and asked for their name and ID.

The Identify mode is used in police investigations to compare a single input image (such as a CCTV image) against a database to generate a candidate list of the most similar images (N=200). The report emphasizes that the algorithm is used as a tool to assist the officers in their investigations. To help clarify the algorithm's role in investigations, the authors refer to the system as 'Assisted Face Recognition' rather than 'Automated Face Recognition'.

Algorithms search databases far faster than the human eye, making it possible to compare an input image against a large number of potential matches. This increases the likelihood that a correct match will be returned. However, the latest algorithms may return multiple faces that are of high visual similarity to the target, and to each other. The increase in algorithm accuracy has resulted in a challenging image comparison task for the human operator when selecting the target from the returned candidate list (see White, Dunne et al., 2015).

Section 3. Review of Face Recognition Accuracy

Automatic face recognition systems are most useful if they increase not only the speed of face recognition, but also the accuracy of facial identifications above that of human performance. This section begins with a brief review of the strengths and limitations of human face recognition and is followed by a review of algorithm face recognition performance from the early 2000s to present day.

Human face recognition accuracy: the alternative to algorithms.

Prior to the invention of automatic face recognition systems, human operators controlled all levels of the face identification procedure. Nowadays, verification can be performed by humans and by algorithms. In many countries, the prevalence of face recognition technology in policing and security scenarios has increased. However, the use of face recognition technology has been met by recent backlash of privacy concerns (e.g., the San Francisco Face Recognition Ban, https://www.bbc.co.uk/news/technology-48276660). The choice to remove machines from the identification process results in the use of the only alternative to algorithm-human systems—an entirely human based system. Before the identification accuracy of algorithms is reviewed, human face recognition performance must be considered. Human performance provides a baseline against which algorithm accuracy can be compared.

Humans are extremely accurate at recognising the faces of people who they know well (e.g., the faces of friends and family members). Identification and matching scenarios are much more challenging when the face is unknown (unfamiliar) to the viewer (Burton, White, & McNeill, 2010; Jenkins, White, Van Montfort, & Burton, 2011). In unfamiliar face matching tasks, humans make errors on approximately 20% of trials under optimized conditions (Burton, Wilson, Cowan, & Bruce, 1999; Burton et al., 2010). Training does little to improve performance (Dowsett & Burton, 2014; Towler, A., White, D. & Kemp, 2017; Towler, White, & Kemp, 2014; Towler et al. 2019, White, Kemp, Jenkins, & Burton, 2014; Woodhead, Baddeley, & Simmonds, 1979), and passport officers perform with comparable accuracy to undergraduate students on these tasks (White, Kemp, Jenkins, Matheson, & Burton, 2014). There are, however, some professional groups, such as forensic facial examiners with access to the tools used for casework (Phillips et al., 2018) and specialist passport officers (White, Phillips, et al., 2015), who as a group, outperform student control groups.

There are large individual differences in face recognition ability. The ability falls on a normal distribution spectrum. One method to increase face recognition performance in security contexts is to recruit people from the top end of this distribution (often referred to as ‘super-recognisers’) into jobs where this skill is important (Noyes, Phillips, & O’Toole, 2017). However, the consistency of performance of super-recognisers across time, and across tasks, is yet to be established (see Bate, Mestry, & Portch, this volume; see also Bate, Portch, Mestry, & Bennetts, 2019; Ramon, Bobak, & White, 2019; Young & Noyes, 2019)

Human face recognition performance is impaired by several factors (see Fysh, this volume). These include low image quality (Bindemann, Attard, Leach, & Johnston, 2013; Burton et al., 1999) and differences in image properties between the target and comparison images (e.g., pose, illumination, camera-to-subject distance) (Hill & Bruce, 1996; Noyes & Jenkins, 2017; O’Toole, Edelman, & Bülthoff, 1998). Human face recognition performance deteriorates when conducting face recognition tasks for long periods of time (Fysh & Bindemann, 2017). Human face recognition is also subject to race bias, known as the ‘other-race effect’ (Feingold, 1914; Malpass & Kravitz, 1969). A predominant theory to explain the other-race effect is the experience or ‘contact’ hypothesis (Carroo, 2011; Chiroro & Valentine, 1995). According to the experience model, human face recognition ability is fine-tuned to the

faces with which we have most experience during childhood (cf., Kelly et al., 2005, 2007).

In sum, even people who are very skilled at unfamiliar face identification, make errors on these tasks. Human performance is impaired by factors such as image quality, fatigue, and the other race effect (Burton et al., 1999; Fysh & Bindemann, 2017; Kelly et al., 2007). Moreover, given the large volumes of digital image evidence generated in criminal investigations via CCTV, social media, and smartphone cameras, there is an increasing demand by police agencies for tools that use this evidence effectively. Humans are simply unable to process these images in large volumes. Algorithms make it possible to make use of this data as they can rapidly compute identity decisions for large databases of faces.

Assessing Algorithm Accuracy

Algorithms address several human shortcomings in face recognition. They can process large databases quickly, and are immune to some factors that effect human error such as fatigue. But how accurate are these algorithms? And are they really better than humans?

There are several challenges associated with evaluating the accuracy of face recognition algorithms in a general and objective way. Meaningful comparisons across algorithms require comparisons to be made on the same task, and on the same image database. In the 1990s, the US Government sponsored competitions to measure algorithm accuracy on the same tasks and images. Data from these competitions now spans decades and demonstrates large improvements in algorithm accuracy across time (Phillips et al., 2005, 2012; Phillips, Moon, Rizvi, & Rauss, 2000). Recent years have seen an increase in formalized objective comparisons across machines, and between humans and machines (Huang, Ramesh, Berg, & Learned-Miller, 2008; Kemelmacher-Shlizerman, Seitz, Miller, & Brossard, 2016; Phillips et al., 2018; Phillips, Hill, Swindle, & O'Toole, 2015; O'Toole, An, Dunlop, & Natu, 2012). The Government sponsored competitions have typically attracted commercial competitors. University-led competitions have also become popular among a largely academic audience (e.g., Huang et al., 2008; Kemelmacher-Shlizerman et al., 2016)

In these competitions, tasks can include 1:1 image matching, 1:N image matching, or image classification. The databases used in these competitions have changed across time to reflect improvements in camera quality, and also to encompass greater image variability. Early databases consisted of frontal images (Phillips et al., 2000), whereas later datasets introduced factors such as varied illumination, pose, 3D images (Phillips et al., 2005), and uncontrolled images (e.g., 'images in-the-wild') (Huang et al., 2008). Such competitions have led to the creation of highly challenging image databases (e.g., IJB-C, see Maze et al. 2018).

Early Algorithms

The algorithms that competed in the early government challenges (Phillips et al., 2000; 2003; 2005) fall into the category of 'early algorithms' that we described in Section 1. In 2005, the state of the art for these algorithms performed was high accuracy on highly controlled frontal-to-frontal image matching tasks, but these algorithms did not fare well on unconstrained images. Algorithm performance

progressed over time, and later versions of these early algorithms rivalled humans on one-to-one image comparison tasks that involved highly controlled frontally-posed datasets (Jenkins & Burton, 2008; O'Toole et al., 2007; 2012). Notably, the accuracy of these algorithms was greatly reduced for face images that varied in pose, illumination, or expression (Phillips et al., 2012, 2015; Sankaranarayanan, Alavi, Castillo, & Chellappa, 2016; Sengupta et al., 2016; O'Toole et al., 2012). By 2012, some early algorithms were comparable to, or better than humans at matching frontal images across changes in illumination (indoor versus outdoor lighting) (O'Toole et al., 2012). However, this was the limit of their capability. Several more recent studies have tested algorithms of these early designs on more natural image types, known as 'in the wild' images. Early algorithm accuracy is far lower for these uncontrolled images than for controlled images (Sankaranarayanan et al., 2016; Sengupta et al., 2016).

State of the Art Performance

Whereas the effective operation of early algorithm performance is limited to good quality, front-facing images, newer algorithms based on DCNNs perform well across a range of image scenarios (Taigman, Yang, & Ranzato, 2014). This is reflective of the varied image sets that they are trained on (see Section 1). Ranjan et al. (2018) reports face identification results for highly challenging datasets (IJB-A, -B, and -C), and Challenge Set 5 (most recent). The galleries in Ranjan et al. (2018) each consisted of over 1 million 'in the wild' face images. Results are reported for various false accept cut off points. However, here we report results at the 0.1% false accept rate, as this is consistent with the false accept rate for ABC e-gates discussed in Section 2.

The challenges spanned two types of task. These were 1:1 verification tasks in which the algorithm decided if two probe images were of the same identity, and a 1:N mixed search, sometimes referred to as a 1:many task. In 1:N tasks, the algorithm receives a probe (input) image, and creates a candidate list of N possible matches, ranked in order of similarity to the probe.

The top performing algorithm performed with over 94% accuracy on all three versions (IJB-A, -B, and -C) of the 1:1 verification tasks at the 0.1% false accept rate. It also performed at over 98% accuracy for this task on the Challenge Set 5 dataset. On the 1:N task, this same algorithm returned the correct rank 1 candidate over 95% of the time in all but one of six tasks (accuracy was at 90.8% in the exception case). For the latest dataset (Challenge Set 5), accuracy rose to 96.99%. Across all datasets, the algorithm returned the correct match within the top 10 rank candidates over 98% of the time. These results demonstrate that the latest algorithms can achieve very high levels of performance accuracy on uncontrolled images.

Fusion

Accuracy was amplified by combining the results of *different* algorithms through fusion (Ranjan et al., 2018). This process mirrors that of a phenomenon known as 'wisdom of the crowds', which is when the collective opinion is more accurate than the opinion of the individual. This effect has been observed for fusing identity judgments in humans (Jeckeln, Hahn, Noyes, Cavazos, & O'Toole, 2018; White et al., 2014).

Link Between Size of Training Database and Performance

Zhou, Cao, and Yin (2015) report a noteworthy pattern in the data on DCNN face recognition: the link between the size of the training database and the performance of the DCNN. As noted in Section 1, DCNNs are trained on a staggering amount of images (millions of images consisting of thousands of identities). Increasing the size of the training dataset appears to increase the performance of the algorithm. In other words, an algorithm performs better if it has been trained on more data.

Three conclusions can be made for DCNN performance, 1) state-of-the-art algorithms perform with very high accuracy on tasks that involve 'in the wild' faces, 2) fusing similarity scores of multiple algorithms can increase identification accuracy, and 3) increasing the size of the training data increases DCNN performance.

Algorithm Bias

It is important to note that algorithms demonstrate several patterns of image bias. For example, identification accuracy of male faces is often higher than identification accuracy for female faces (e.g., Blanton, Allen, Miller, Kalka, & Jain, 2016). Additionally, Phillips et al. (2011) report that an 'other-race effect' for both humans and early algorithms. In their study, half of the human participants were Caucasian, and the other half was East Asian. A Caucasian algorithm (made by fusing similarity scores of 8 algorithms made in Western countries) and an East Asian algorithm (made by fusing 5 algorithms created in East Asian countries) were also tested. Performance was measured on recognition of highly controlled images of Caucasian and East Asian Faces. Humans performed more accurately on faces of their own race. Algorithms performed best on the predominant race where the algorithm was developed (Phillips et al., 2011).

These published findings are based on older class algorithms. However bias has also been reported for DCNNs (Khiyari & Wechsler, 2016; Krishnapriya, Vangara, King, Albiero, & Bowyer, 2019). This may reflect bias in training data (Klare, Burge, Klontz, Vorder Bruegge, & Jain, 2012), and/or differences in image quality (Krishnapriya et al., 2019). A very recent assessment of algorithm performance from the Face Recognition Vendor Test reports that the four top performing algorithms made least errors for black male faces when compared against performance for male white, female white, and female black faces (Klare, 2019). Demographic differentials were observed in the majority of algorithms tested in the NIST Face Recognition Vendor Test (Grother, Ngan & Hanaoka, 2019). The extent of these differences depended upon both the algorithm, and the task at hand. It is therefore critical that users know the capabilities and limitations of their algorithm and set appropriate criterion thresholds to minimise bias.

Comparison of human and algorithm accuracy

Algorithm and human face recognition is a topic that spans both the psychology and the computer science literature. Several computer science papers include data on human face recognition accuracy. However, the methods used to assess human performance in the computer science literature are often inconsistent with those used in the psychology literature. The following points must be kept in mind when making conclusions on human and machine performance from the computer science literature. Psychologists tend to compute the average accuracy of each participant as derived

from their individual (i.e. non-aggregated) response. O'Toole and Phillips (2017) argue that computer science reports of human face recognition accuracy are often inflated through misuse of fusion. Instead of averaging performance of human participants over a task, the classification accuracy of humans is often derived by averaging human judgments to pairs of faces. This inflates accuracy due to a phenomenon known as 'wisdom of the crowds'. Computer science papers often report this fused human score, when the *average* human score should be reported. Computer science studies also often fail to account for factors such as face familiarity, the other-race effect, and access to information other than the face (e.g., body or clothing), all of which can influence human performance (O'Toole & Phillips, 2017).

O'Toole et al. (2012) provide the first direct comparison of human and algorithm 1:1 face-matching performance. In their study, humans (undergraduate students) and algorithms were tested on their ability to match a subset of images from the Face Recognition Grand Challenge image set. The images were carefully selected to contain 'easy' (constant illumination) and 'difficult' (varied illumination) image pairs. Algorithms consistently outperformed these untrained human lay observers on the easy image pairs. Moreover, three out of the four algorithms that were tested also outperformed humans on the difficult image pairs (O'Toole et al. 2012).

At the beginning of Section 3 we noted that some individuals are better at recognising faces than others. Super-recognisers and forensic examiners with access to their tools perform particularly well on face recognition tasks (see White, Towler, & Kemp, this volume). Phillips et al. (2018) compare the performance of four DCNNs (state-of-the art in 2015, through to 2017) against undergraduate students, super-recognisers, and forensic examiners with access to their tools on a difficult face-matching task involving 20 frontal-facing face image pairs. The images varied in illumination and—to a lesser degree—in expression. The most recent algorithm (2017b) performed with accuracy levels that were comparable to the top performing human participant groups (the forensic examiners and super-recognisers). Algorithms have made dramatic accuracy gains between 2015 and 2017. The 2015 algorithm performed with accuracy levels that are comparable to the score of the median undergraduate student, whereas the 2017b algorithm performed at a comparable level to the median forensic examiner. This means that some examiners were better than the algorithm and some were worse, but the algorithm was equal to the 'middle value' of forensic examiners.

Accuracy of Algorithms out in the Field

Thus far, we have covered algorithm accuracy as tested in controlled experimental settings. If these algorithms are to be used in real-world applications, it is important to test accuracy of algorithms 'in the field'. Davies, Innes, and Dawson (2018) provide a report of South Wales Police Usage of Automated Face Recognition Systems. Their report explains two modes of usage of algorithms in police investigations. These are termed 'Locate' and 'Identify' (explained in Section 2).

A field test was designed to test accuracy of the 'Locate' system. In this test, images of police officers were added to a watch list, and the system was tested on how accurately it flagged these officers when their live face image was captured by the system. The ground truth (true identity) of the officers was known, making it possible to assess algorithm accuracy. The algorithm flagged a true match between the officer

and image on the watch list in between 76% and 81% of trials. In practice, a human officer would be required to review and verify the match.

The report also includes accuracy of the Locate system in operational deployments. The database images were 1200 images, made up of traditional police custody 'mug shot' images and other 'non-custody' images, which had been taken outside. Despite the relatively high accuracy of the system in the field test reported above, in applied practice the system flagged many faces, deemed as false positives by the human operator. In initial usage of the algorithm, just 3% of flagged images were considered to be a true match. However, this number rose to 46% in later deployments (Davies, Innes, & Dawson, 2018). Arguably, the remaining 54% of identifications cannot be considered as false positives because the human reviewer will ultimately decide whether to follow up the flagged identity to make a positive identification. Notwithstanding, the algorithm deployment resulted in a small number of arrests.

The report outlines several challenges experienced with the Locate system. Namely, image quality, lighting, occluded features, and operational issues. The algorithm required good quality database images in order to compute accurate similarity scores. Lighting affected matching accuracy, with the system performing poorly in dim lighting. Additionally, certain clothing and accessories, which obscured parts of the face when worn, reduced algorithm accuracy. And in live matching scenarios, the system often lagged, froze, or crashed, when dealing with images of crowds (Davies, Innes, & Dawson, 2018).

The success of the Identify system was also evaluated. Here, database images consisted of around 45,000 police custody images (mug shot style images). The exact number increased over time. The quality of the probe image had a substantial effect on system accuracy. Between the end of October 2017 and March 2018, the algorithm rejected 60% of all input images due to poor image quality. Many of these rejected images had were mobile phone images of CCTV footage. In this same time frame, 73% of the images that were accepted by the algorithm returned a possible true match within the candidate list, as confirmed by a police operator. Within these cases, the true match was listed as the rank 1 candidate 60% of the time, and listed within the top 10 ranked candidates 90% of the time. Notably, the true match here cannot necessarily be considered a ground truth as humans also make face matching errors.

Section 4. Interactions between algorithms and humans in face recognition

In this section we consider the advantages and disadvantages of interactions between algorithms and humans. Specifically, we discuss the benefits of combining the response of algorithms and humans through fusion, and also consider the influence that an algorithm's identification may have on the human decision making process.

Fusion of Algorithm and Human Similarity Scores

It is well documented that accuracy of human face recognition can be improved by wisdom of the crowds—fusing the decisions of multiple people on an item-by-item basis, or simply put, the combined judgment of many is better than the decision of an individual (Jeckeln, Hahn, Noyes, Cavazos, & O'Toole, 2018; White, Burton, Kemp, & Jenkins, 2013). There is a similar benefit for fusing the performance of multiple algorithms (Ranjan et al., 2018).

Fusing the response of humans with that of algorithms has led to large performance gains over the response of either humans or algorithms alone (O'Toole, Abdi, Jiang, & Phillips, 2007; Phillips et al., 2018). O'Toole, Abdi, et al. (2007) report that fusion of human and algorithm similarity scores resulted in near perfect accuracy on the task outlined in Section 3. Additionally, Phillips et al. (2018) report that highest identification accuracy was achieved by combining the decisions of the top performing humans (forensic examiners) with the identifications provided by the algorithm. Fusion of human and machine identification decisions works in boosting overall performance, because humans and algorithms most likely use different methods to compute their similarity calculations. For example, some images that are challenging to an algorithm are not similarly challenging to humans, and vice versa (O'Toole et al, 2012). Fusion utilises the individual strengths of both humans and algorithms.

Interference of systems

In applied face identification scenarios that involve both humans and machines, humans act as check and balance against algorithm identification. In an ideal operational scenario, the human and algorithm will agree on correct identification verdicts. When the algorithm is incorrect, the human must catch and correct the algorithm error. In practice, the human operator may receive the algorithm output (similarity score) prior to reaching their identification decision. This raises an important question of whether the human operator is influenced by the algorithm's verdict.

Fysh and Bindemann (2018) tested whether human face matching is influenced by the presence of a pre-assigned identification label, such as that provided by an algorithm. In their study, participants viewed pairs of face images and made same/different identity responses to each image pair. Each pair of faces had been assigned a label of 'same identity', 'different identity', or 'unresolved' (representing no answer from the algorithm). Participants were told that the assigned label would often be accurate; this was true and reflects the high accuracy rates of current algorithms. The result was that the label influenced the identification; incorrect identifications were made most often for image pairs that had an incorrect label. Instruction to ignore the label made no difference to the label influence. Performance on the trials labelled 'unresolved' was in line with performance on standardised face matching tasks (e.g., Burton et al. 2010). Additionally, Heyer, Semmler and Hendrickson (2019) report that candidate list length affects accuracy of a human reviewer Candidate lists of over 100 items produced more false alarms, fewer hits, and lower confidence in identifications than smaller candidate lists. Further testing is necessary to assess the effect of algorithm output on the accuracy of the human reviewer in practice.

When humans are the weak link in the system.

The accuracy of any human-algorithm system is limited by the accuracy of each of these components—the human and the algorithm. If the human operator holds authority over the final identification, then the algorithm's performance is capped by the accuracy of the human operator.

Algorithms regularly return a true match to an input image within a candidate list, however, it is up to a human to select and confirm a match from this list. White,

Dunn, Schmid, and Kemp (2015) argue that humans may be the weak link in this process. They tested accuracy of undergraduate students and passport officers for selecting a match, or identifying a target as absent from a candidate list generated by an algorithm. In White et al. (2015), participants made errors in selecting the correct match from a candidate list on 50% of trials. Furthermore, there was no difference in performance of undergraduate students and passport officers. However, a specialist group of face examiners made 20% less errors. The results from White et al. (2015) demonstrate that even if algorithms are highly accurate at returning a correct match within a candidate list, humans are not always accurate at selecting the match from the list.

Section 5. Summary

State-of-the-art face recognition algorithms modeled on DCNNs are far more accurate than their predecessors, and can operate accurately over greater image variation. This is because DCNNs leverage deep architectures and are often trained with millions of images of thousands of identities. Between the years of 2015 to 2017, algorithm face-matching accuracy increased from that of the median undergraduate student, to that of the median forensic examiner (Phillips et al., 2018). It is likely that DCNN performance will continue to rise as training datasets become larger and algorithm developers begin to tackle more challenging image scenarios.

Algorithm accuracy is increasing at such a rapid rate that the literature that reports state of the art performance quickly becomes outdated. The latest algorithms perform with very high accuracy on image-matching tasks involving front facing images (Phillips et al., 2018), and also on much more challenging, naturalistic images (Ranjan et al., 2018). Ranjan et al. (2018) report impressive return rates of true matches within top ten (and increasingly rank 1) images on candidate lists.

How do algorithms compare with humans? Algorithms perform more accurately than the average human on many frontal, 1:1 and 1:N, image-matching tasks. Algorithms and humans performance on images in the wild has not yet been compared directly, however we know that this is often a difficult task for humans, and algorithms are scoring with increasingly high accuracy on these types of tasks. In terms of speed and breadth of search scope, algorithms far outweigh humans. Algorithms can search databases of millions of images and return a list of possible matches within seconds. Despite the impressive accuracy rates of algorithms, human verification remains an important part of the face identification process. In operational settings, humans are most often required to inspect and review the algorithm output. Face recognition algorithms have been described as a tool that can be used to assist investigations (Davies et al., 2018).

Whilst algorithms have several strengths, it is also important to consider their limitations. Studies have revealed both gender and race biases in algorithm face identification (Blanton et al., 2016; Phillips et al., 2009). It is important to remember that humans also exhibit an other-race effect in their face recognition performance; typically, performance is higher for recognizing faces of own than other races (Carroo, 2011; Malpass & Kravitz, 1969). The experience hypothesis explains this in terms of fine-tuning face recognition expertise to races experienced most during childhood (Kelly et al., 2005, 2007). Analogously, algorithm race bias might be

linked to an imbalance in faces of certain races represented within training data (Klare et al., 2012).

Algorithms also experience several challenges in live deployment scenarios. Many of the issues with live deployment have been liked to poor image quality. Additionally, large number of false positive responses result in a large workload for the human operator (Davies et al., 2018). Lighting and clothing can also affect algorithm accuracy (Davies et al., 2018). Despite these issues, the use of algorithms in live deployment scenarios has resulted in a small number of arrests. It is important to also investigate more deliberate attempts to deceive the system, including masks (Sanders et al., 2017), morphs (Robertson, Kramer, & Burton, 2017), and deliberate disguise (Noyes & Jenkins, 2019).

Face recognition systems rely on the accuracy and efficiency of both humans and algorithms. At times, the accuracy of one may be capped by accuracy of the other (Fysh & Bindemann, 2018; White, Dunn, et al., 2015), however both humans and machines can contribute to the identification effort. Indeed, the best systems may result from fusing the judgments of the best-performing humans with the scores of algorithms. This resulted in the most accurate performance in a recent face matching test for frontal images (Phillips et al., 2018). It is not yet known whether the benefit extends to more challenging images.

Accurate algorithm performance is dependent upon well-considered policy for the use of machines in face identification. The scientific evidence supports the use of face recognition algorithms for front-facing images, and also for naturalistic images that vary in pose, illumination, expression, etc. Each individual algorithm has its own strengths and limitations, as do humans. As these technologies continue to evolve and grow, it is important to understand their strengths and weaknesses to ensure the appropriate use of these algorithms. Each algorithm is different. Operators need to know their algorithm in order to understand its capabilities and to set appropriate thresholds to reduce bias. The science should drive the use of face recognition algorithms and their role in human-machine face identification systems.

Conclusions

Going forward, as machines become more accurate and more integrated into our daily lives, there will be important questions to consider. If algorithms consistently outperform humans, then what role should humans play in the face recognition process? Perhaps the role of the human will change from that of an equal partner to the algorithm, to that of ‘error catcher’, or a system manager who knows the capabilities of the algorithm and sets appropriate parameters. For example, does the image quality meet the requirements for accurate identification? Is the criterion threshold appropriate for the demographics? If humans and algorithms perform identifications in different ways, then there will always be a role for humans to catch the errors made by machines. We expect that the most accurate identification systems will include a role for top performing humans and top performing algorithms.

Acknowledgements

This work was supported by the NIH under grant R01EY029692. Thank you to Professor Alice J. O'Toole for helpful comments on a draft of this chapter.

References

Alenezi, H. M., Bindemann, M., Fysh, M. C., & Johnston, R. A. (2015). Face matching in a long task: enforced rest and desk-switching cannot maintain identification accuracy. *PeerJ*, *3*, e1184. https://doi.org/10.7717/peerj.1184

Bate, S., Portch, E., Mestry, N., & Bennetts, R. J. (2019). Redefining super recognition in the real world: Skilled face or person identity recognizers? *British Journal of Psychology*, 2017–2019. https://doi.org/10.1111/bjop.12392

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*(10). https://doi.org/10.1098/rsos.160321

Bindemann, M., Attard, J., Leach, A. M. Y., & Johnston, R. A. (2013). The Effect of Image Pixelation on Unfamiliar-Face Matching, *717*(November), 707–717.

Blanton, A., Allen, K. C., Miller, T., Kalka, N. D., & Jain, A. K. (2016). A Comparison of Human and Automated Face Verification Accuracy on Unconstrained Image Sets. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 229–236. https://doi.org/10.1109/CVPRW.2016.35

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence From Security Surveillance. *Psychological Science*, *10*(3), 243–248. https://doi.org/10.1111/1467-9280.00144

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*, 286–291.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12). https://doi.org/10.1371/journal.pcbi.1003963

Carroo, A. W. (2011). Other Race Recognition: A Comparison of Black American and African Subjects. *Perceptual and Motor Skills*, *62*(1), 135–138. https://doi.org/10.2466/pms.1986.62.1.135

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, *48*(4), 879-894. https://doi.org/10.1163/_q3_SIM_00374

Davies, B., Innes, M., & Dawson, A. (2018). An Evaluation of South Wales Police's Use of Autormated Facial Recognition. *Universities' Police Science Institute Crime and Security Research Institute. Report*. Retrieved from

https://www.statewatch.org/news/2018/nov/uk-south-wales-police-facial-recognition-cardiff-uni-eval-11-18.pdf

Dowsett, A. J., & Burton, A. M. (2014). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*, 433–445.

Feingold, G. A. (1914). The Influence of Environment on Identification of Persons and Things. *Journal of the American Institute of Criminal Law and Criminology*, *5*(1), 39. https://doi.org/10.2307/1133283

FRONTEX. (2015). *Best Practice Technical Guidelines for Automated Border Control (ABC) Systems*. Retrieved from https://books.google.co.il/books?id=bYONnQAACAAJ

Fysh, M. C., & Bindemann, M. (2017). Effects of time pressure and time passage on face-matching accuracy. *Royal Society open science*, *4*(6), 170249.

Fysh, M. C., & Bindemann, M. (2018). Human–Computer Interaction in Face Matching. *Cognitive Science*, *42*(5), 1714–1732. https://doi.org/10.1111/cogs.12633

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3 : Demographic Effects. *NISTIR 8280*, *December*, https://doi.org/10.6028/NIST.IR.8280.

Heyer, R., Semmler, C., & Hendrickson, A. T. (2018). Humans and Algorithms for Facial Recognition: The Effects of Candidate List Length and Experience on Performance. *Journal of applied research in memory and cognition*, *7*(4), 597-609.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(4), 986. https://doi.org/10.1037/0096-1523.22.4.986

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2008). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.*Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 1–11.

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., & O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, 1–12. https://doi.org/10.1111/bjop.12291

Jenkins, R., & Burton, A. M. (2008). 100% Accuracy in automatic face recognition. *Science*, *319*(5862), 435. https://doi.org/10.1126/science.1149656

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

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007a). The Other-Race Effect Develops During Infancy. *Psychological Science*, *18*(12), 1084–

1089. https://doi.org/10.1111/j.1467-9280.2007.02029.x

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., & Pascalis, O. (2007b). The Other-Race Effect Develops During Infancy Evidence of Perceptual Narrowing, *18*(12), 1084–1089.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., Ge, L. & Pascalis, O. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*(6), 31–36.

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The MegaFace benchmark: 1 million faces for recognition at scale. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 4873–4882. https://doi.org/10.1109/CVPR.2016.527

Khiyari, H. El, & Wechsler, H. (2016). Journal of Biometrics & Biostatistics Face Verification Subject to Varying (Age , Ethnicity , and Gender) Demographics Using Deep Learning. *Journal of Biometrics & Biostatistics*, *7*(4). https://doi.org/10.4172/2155-6180.10003

Klare, B.F., (2019). https://blog.rankone.io/2019/09/12/race-and-face-recognition-accuracy-common-misconceptions/

Klare, B. F., Burge, M. J., Klontz, J. C., Vorder Bruegge, R. W., & Jain, A. K. (2012). Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, *7*(6), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

Krishnapriya, K. S., Vangara, K., King, M. C., Albiero, V., & Bowyer, K. (2019). Characterizing the Variability in Face Recognition Accuracy Relative to Race. *ArXiv*, *arXiv:1904*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105.

Maze, B., Adams, J.C., Duncan, J.A., Kalka, N.D., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., & Grother, P. (2018). IARPA Janus Benchmark - C: Face Dataset and Protocol. *2018 International Conference on Biometrics (ICB)*, 158-165

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of personality and social psychology*, *13*(4), 330.

Nech, A., & Kemelmacher-Shlizerman, I. (2017). Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7044-7053.

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

Cultural Differences (pp. 173–201). New York: Nova Science Publishers, Inc.

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition, 165*, 97–104. https://doi.org/10.1016/j.cognition.2017.05.012

Noyes, E. & Jenkins, R. (2019). Deliberate disguise in face identification . *Journal of Experimental Psychology Applied, 25*(2), 280-290. http://dx.doi.org/10.1037/xap0000213

O'Toole, A. J., Abdi, H., Jiang, F., & Phillips, P. J. (2007). Fusing face-verification algorithms and humans. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 37*, 1149-1155.

O'Toole, A. J., An, X., Dunlop, J., & Natu, V. (2012). Comparing Face Recognition Algorithms to Humans on Challenging Tasks. *ACM Transactions on Applied Perception (TAP), 9*(4), 1–13. https://doi.org/10.1145/2355598.2355599

O'Toole, A. J., Edelman, S., & Bülthoff, H. H. (1998). Stimulus-specific effects in face recognition over changes in viewpoint. *Vision Research, 38*, 2351–2363.

O'Toole, A. J., Phillips, P. J., Member, S., Jiang, F., Ayyad, J., & Pe, N. (2007). Face Recognition Algorithms Surpass Humans Matching Faces over Changes in Illumination, *29*(9), 1642–1646.

O'Toole, A. J., & Phillips, P. J. (2017). Five Principles for Crowd-Source Experiments in Face Recognition. *Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition,* 735–741. https://doi.org/10.1109/FG.2017.146

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D., … Weimer, S. (2012). The good, the bad, and the ugly face challenge problem. *Image and Vision Computing, 30*(3), 177–185. https://doi.org/10.1016/j.imavis.2012.01.004

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., … Worek, W. (2005). Overview of the face recognition grand challenge. *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 947–954. https://doi.org/10.1109/CVPR.2005.268

Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., & Bone, M. (2003). Face recognition vendor test 2002. *IEEE International SOI Conference. Proceedings*, 44. https://doi.org/10.1163/_q3_SIM_00374

Phillips, P.J., Hill, M. Q., Swindle, J. A., & O'Toole, A. J. (2015). Human and algorithm performance on the PaSC face Recognition Challenge. *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems. 1-8. https*://doi.org/10.1109/BTAS.2015.7358765

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, *8*(2), 14.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(10), 1090–1104.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, JC, Castillo, C. D., Chellappa, R., White, D. & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171–6176. https://doi.org/10.1073/pnas.1721355115

Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. *British Journal of Psychology*, 1–19. https://doi.org/10.1111/bjop.12368

Ranjan, R., Bansal, A., Zheng, J., Xu, H., Gleason, J., Lu, B. Nanduri, A., Chen, JC., Castillo, C. & Chellappa, R. (2018). A Fast and Accurate System for Face Detection, Identification, and Verification, *14*(8), 1–16. Retrieved from http://arxiv.org/abs/1809.07586

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2017). Fraudulent ID using face morphs : Experiments on human and automatic recognition, 1–12. https://doi.org/10.1371/journal.pone.0173319

Sanders, J. G., Ueda, Y., Minemoto, K., Noyes, E., Yoshikawa, S., & Jenkins, R. (2017). Hyper-realistic face masks : a new challenge in person identification. *Cognitive Research*, *2*(43). https://doi.org/10.1186/s41235-017-0079-y

Sankaranarayanan, S., Alavi, A., Castillo, C. D., & Chellappa, R. (2016). Triplet probabilistic embedding for face verification and clustering. *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems, BTAS 2016*, 1–8. https://doi.org/10.1109/BTAS.2016.7791205

Sengupta, S., Chen, J. C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016). Frontal to profile face verification in the wild. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, 1–9. https://doi.org/10.1109/WACV.2016.7477558

Taigman, Y., Yang, M., & Ranzato, M. A. (2014). Deepface: Closing the gap to human -level performance in face verification. *CVPR IEEE Conference*, 1701–1708. https://doi.org/10.1109/CVPR.2014.220

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, *3*(1), 71-86.

Towler, A., White, D. & Kemp, R. (2017). Evaluating the feature comparison

strategy. *Journal of Experimental Psychology: Applied*, *23*, 47.

Towler, A., White, D., & Kemp, R. I. (2014). Evaluating training methods for facial image comparison: The face shape strategy does not work. *Perception*, *43*(2–3), 214–218. https://doi.org/10.1068/p7676

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work ? *PLoS ONE*, *14*(2), e0211037. https://doi.org/10.1371/journal.pone.0211037

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. O. B. (2013). Crowd Effects in Unfamiliar Face Matching, *27*, 769–777.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE*, *10*(10), 1–14. https://doi.org/10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., & Burton, a M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review*, *21*, 100–106.

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., & White, D. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20151292.

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

Young, A. W., & Noyes, E. (2019). We need to talk about super-recognizers : Invited commentary on: Ramon, M., Bobak, A. K., & White, D. Super-recognizers: From the lab to the world and back again. British Journal of Psychology. *British Journal of Psychology*, 2017–2019. https://doi.org/10.1111/bjop.12395

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*. 818-833.

Zhou, E., Cao, Z., & Yin, Q. (2015). Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv preprint arXiv:1501.04690*.