Perceptual and computational detection of face morphing

Sophie J. Nightingale School of Information, University of California, Berkeley, CA, USA

Shruti Agarwal Electrical Engineering & Computer Sciences, University of California, Berkeley, CA, USA

Hany Farid Electrical Engineering & Computer Sciences and School of Information, University of California, Berkeley, CA, USA

A relatively new type of identity theft uses morphed facial images in identification documents in which images of two individuals are digitally blended to create an image that maintains a likeness to each of the original identities. We created a set of high-quality digital morphs from passport-style photos for a diverse set of people across gender, race, and age. We then examine people's ability to detect facial morphing both in terms of determining if two side-by-side faces are of the same individual or not and in terms of identifying if a face is the result of digital morphing. We show that human participants struggle at both tasks. Even modern machine-learning-based facial recognition struggles to distinguish between an individual and their morphed version. We conclude with a hopeful note, describing a computational technique that holds some promise in recognizing that one facial image is a morphed version of another.

Introduction

We frequently rely on photo-based identity documents to verify identity in critical settings such as border control. Much research, however, has shown that matching pairs of unfamiliar faces is a difficult task (Bruce et al., 1999; Megreya & Burton, 2007; Burton et al., 2010), including for trained identification-checkers (White et al., 2014). The difficulty of this task leaves identity verification processes vulnerable to fraudulent attacks. One relatively new type of fraud that border control agencies are facing is the use of morphed passport photos. Early research exploring human detection of morphed images indicates that people frequently accept both low-quality and high-quality morphed images as genuine (Robertson et al., 2017; Robertson, 2018; Kramer et al., 2019). The face databases used in these previous studies, however, have a number of limitations, most notably low-quality morphs and/or limited diversity in terms of the race, gender, and age of the faces used to create the stimuli. Here, we extend this previous literature by examining human and computer-based detection of face morphing using a diverse set of facial images to generate high-quality morphs.

Attempts to obtain fraudulent identity documents are not new. Traditionally, it was common for fraudsters to attempt to create fraudulent identifications by producing fake passports or removing and replacing the photo in a real passport. In response, passport anti-counterfeit measures were developed, making such attempts easy to detect (UK HM Passport Office, 2020) (e.g., patterns that are visible only under specific artificial illumination). In light of these anti-counterfeit measures, fraudsters have started to find ways to obtain fraudulently obtained but genuine (FOG) passports (ITW Security Division, 2017; Middleton, 2014). These FOG passports are real documents that are issued by an official and can therefore circumvent the previously mentioned anti-counterfeit measures. A common strategy for obtaining a FOG passport is through an accomplice, who holds a genuine passport, submitting a renewal application with the photo of a similar-looking fraudster. Most recently, this issue has intensified through the increased availability of advanced image manipulation software that enables fraudsters to create morphed facial images. Face morphing involves digitally combining images of two (or more) individuals to create a manipulated image that maintains a likeness to each of the original identities. To highlight how this process might work, imagine the following scenario: A fraudster (ID1) who *is not* legally able to obtain a passport morphs a photo of their face with the face of another person who *is* legally able to obtain a passport (ID2). This morphed facial image is submitted alongside ID2's passport application. When checked against the photo stored on file at the passport issuance

Citation: Nightingale, S. J., Agarwal, S., & Farid, H. (2021). Perceptual and computational detection of face morphing. *Journal of Vision*, *21*(3):4, 1–18, https://doi.org/10.1167/jov.21.3.4.

https://doi.org/10.1167/jov.21.3.4 Received May 28, 2020; published March 3, 2021 ISSN 1534-7362

office, the morphed face resembles ID2 closely enough to result in the issuance of a passport. And crucially, the morphed face also resembles ID1 closely enough to allow them to pass through border control. Of course, the use of a morphed facial image rather than an image of a similar-looking accomplice carries a clear benefit for the fraudster: the manipulated image contains some of their facial features. This scenario will, however, fail if the morphed facial image is detected at either the issuance or ID-check stage. An important question, then, concerns whether identification examiners (both human and machine) are able to detect morphed facial images.

Research on this topic is in its infancy, but early studies suggest that people cannot reliably detect morphed faces. In one study, participants were presented with two face images and were asked to indicate whether those faces depicted the same person or not (Robertson et al., 2017). Participants completed 49 trials, seven in which the two faces were of the same person, 7 in which the two faces were of different people, and 35 in which a face was paired with a morph of that face and a different person. People accepted 50/50 face morphs (weighting both original identities equally) as a "match" for the face it had been paired with 68% of the time. In a follow-up, participants were given basic guidance on how to detect morphs and also given an additional response option of "morph." This time, 50/50 face morphs were accepted as "matches" 21% of the time. These results indicate that such morphed faces might provide an opportunity to commit identity fraud.

In another study, Robertson and colleagues (2018) examined the effect of simple training on morph detection. Participants were assigned at random to either a guidance-only or guidance and training group. The guidance consisted of basic information about morphing and tips for detection, for example, to look for a ghost-like outline of another face or another person's hair over the forehead. As well as receiving this guidance, participants in the training group completed a two-alternative forced-choice (2AFC) task where they had to determine which of the faces was a morph and were given accuracy feedback after each of the 20 trials. Both before and after the guidance/training, participants were shown a series of trials each consisting of an array containing 10 faces, half morphs and half original. On each trial, participants indicated which of the faces were morphs. Baseline performance before guidance/training, measured as d' sensitivity, was $d' = 0.96$ for the guidance-only group and $d' = 0.56$ for the guidance and training group. After guidance/training, performance improved significantly to $d' = 2.32$ for the guidance-only group and $d' = 2.69$ for the guidance and training group. These results suggest that performance can be significantly improved through guidance and training.

There are, however, a number of important limitations with these two studies. First, although the morphs were created using advanced morphing software, there was no manual editing stage to remove obvious artifacts that are known to result from the morphing process, such as the outline of another person's hair. In fact, such artifacts were precisely what the authors guided participants to look for to help them to detect morphs. This limitation might artificially inflate the effect of the guidance and training manipulation. Second, the stimuli were created using facial images from the Glasgow Face Matching Test (Burton et al., 2010), which includes mostly White individuals. This lack of racial diversity in the stimuli further limits the extent that the results of these studies are likely to be representative of the detection of morphs in the real world.

In 2019, another research group sought to address the first limitation noted above by replicating the study by Robertson et al. (2018) using higher-quality face morphs (Kramer et al., 2019). Using these higher-quality morphs, they also found that initial detection was poor, but unlike Robertson et al.'s (2018) study, training had no effect on accuracy. The use of the 10-image array paradigm makes it difficult to determine chance performance and is not akin to the process identification checkers use in the real world. In a second experiment, participants saw a single image per trial, half morphs and half original images, and were asked to indicate if the image was a morph or not. Half of the participants saw a number of morph detection tips before completing the task while the other half did not receive this information. Both groups performed poorly and the tips did not reliably improve performance—perhaps, then, tips and guidance about morph detection might only be useful when using low-quality morphs that clearly contain visual morphing artifacts. In a third experiment, participants completed a live face-matching task rather than a computer-based one. For this task, 48 models (44 White) were photographed and paired with a visually similar model (foil). The models approached participants on a university campus and presented either a photo of the (1) model (match), (2) model morphed with the foil individual (50/50 morph), or (3) foil (mismatch). The participants were asked to indicate if they thought the photo was of the model or not. For the match and mismatch conditions, average accuracy was 83% and 84%, but the morph photos were accepted nearly half of the time (49%). The pairing of models to create the morph photos revealed an interesting finding: In the majority of the pairs, the morph was accepted as a valid identification for one model more frequently than for the other model. Kramer et al. (2019) conclude that even when generating 50/50 morphs, the morph does not represent each of the original faces equally. Because in previous work (e.g., Robertson et al., 2017), the morphs were only presented

alongside one of the original identities, the results might not accurately represent true morph detection rates.

These three experiments indicate that human ability to detect morph faces is extremely limited and that basic training is not sufficient to improve this performance. In a final experiment, Kramer et al. (2019) tested whether a simple computational model could outperform human ability to detect morphs. Principal components analysis was used to extract a low-dimensional representation of the faces. These representations were then used to train a linear discriminant analysis model with two classes, morph and original. Using this model to classify the remaining images that were not used in the training set resulted in an average accuracy of 68%, corresponding to a sensitivity of 1.01. This result suggests that a simple computational model can outperform humans at morph detection but remains a far from perfect classifier.

Identity fraud is a serious issue that poses significant risk to national security. In fact, in Germany, the threat of face morphing to commit identity fraud has prompted plans to heighten security by making people take passport photos in official government-owned booths that transfer the photo directly to official computers (Huggler, 2020). Given the significance of the threat face morphing poses, it is critical that further research is conducted to better understand human and machine ability to detect morphing as well as finding reliable ways to improve detection. To advance work on this important topic, we describe the creation of a data set that addresses the limitations noted in previous work in this field. Specifically, our data set includes high-quality morphs that are diverse across race, gender, and age. Addressing limitations noted in Kramer et al. (2019), we selected our image pairs from a large initial pool of faces using a computational approach to automatically find similar-looking faces. In addition, rather than generating 50/50 morphs, we generated morphs that represent each of the original faces equally.

Using this data set, we conduct a series of perceptual experiments examining people's ability to perceptually detect morphed faces. Our approach is rooted in the goal of understanding how accurately an analyst will be able to compare an identity to a morphed identity and how accurately they will be able to recognize a morphed face. This study unfolds in a series of experiments designed to measure accuracy and bias in the corresponding recognition and detection task, as well as examining if certain interventions might improve performance. We conclude with a comparison of perceptual accuracy to strictly computational approaches for detecting morphed faces.

Data set

This section describes the creation of a data set used in all of our experiments. We note four main limitations of other data sets that have been used in similar research (Robertson et al., 2017; Robertson et al., 2018; Kramer et al., 2019): (1) obvious visual morphing artifacts, (2) a small number of faces from which to select matching faces, (3) a manual process to match similar faces, and (4) a lack of racial/gender diversity. We address these limitations to create a high-quality and diverse data set.[1]

We collected 3,500 passport-format facial images from 13 face databases (Phillips et al., 1998, 2000; Flynn et al., 2003; Weyrauch et al., 2004; Phillips et al., 2005; Azam et al., 2007; Kasiński et al., 2008; Utrecht ECVP Face Database, 2008; Wang & Tang, 2008; Thomaz & Giraldi, 2010; Watson, 2010; Vieira et al., 2014; Ma et al., 2015; Strohminger, 2016; DeBruine & Jones, 2017). These 3,500 images included a diverse range of individuals across gender, age, and race. To ensure diversity in our final stimulus set, we manually selected 54 individuals constituting 6 African American or Black, 16 East Asian, 16 South Asian, and 16 Caucasian. Of these, 26 were women and 28 were men, spanning a range of apparent ages. Some of the face databases specified the race/gender of each face. Where this information was not available, we relied on the subjective judgment of the three authors.

We matched each of these 54 individuals with their most similar-looking counterpart in the remaining faces in our data set. A standard convolutional neural network descriptor (termed VGG) (Parkhi et al., 2015) was used to extract a low-dimensional, perceptually meaningful, representation of each face in the full data set. The extracted representation—a 4,096-D real-valued vector—for each of the 54 manually selected target faces was compared with all other representations in the data set to find the most similar face as defined by the face whose representation is most similar—in terms of Euclidean distance—with the target face (Tariq et al., 2018; Zhang et al., 2018). A midway morph was then generated for each pair of matched faces as follows.

A total of 68 corresponding points on the two faces were extracted using a standard facial landmark detector (King, 2009; Figures 1a,b). These points were augmented with an average of 116 manually selected points along the hairline and top of the head, ears, and neck (Figures 1d,e). These manually selected points improved the overall visual quality of the generated morphs (Figure 1c vs. f). After extracting corresponding facial landmarks, and prior to generating the facial morphs, the two faces were aligned by an affine transform, consisting of anisotropic scaling, shearing, rotation, and translation. This alignment ensured that facial features did not significantly move during the morphing process. In particular, denote the 68 corresponding feature points on each face as (x_i, y_i) and (u_i, v_i), $i \in [1, 68]$. The six-parameter affine is given by

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix} + \begin{pmatrix} a_5 \\ a_6 \end{pmatrix}, \quad (1)$$

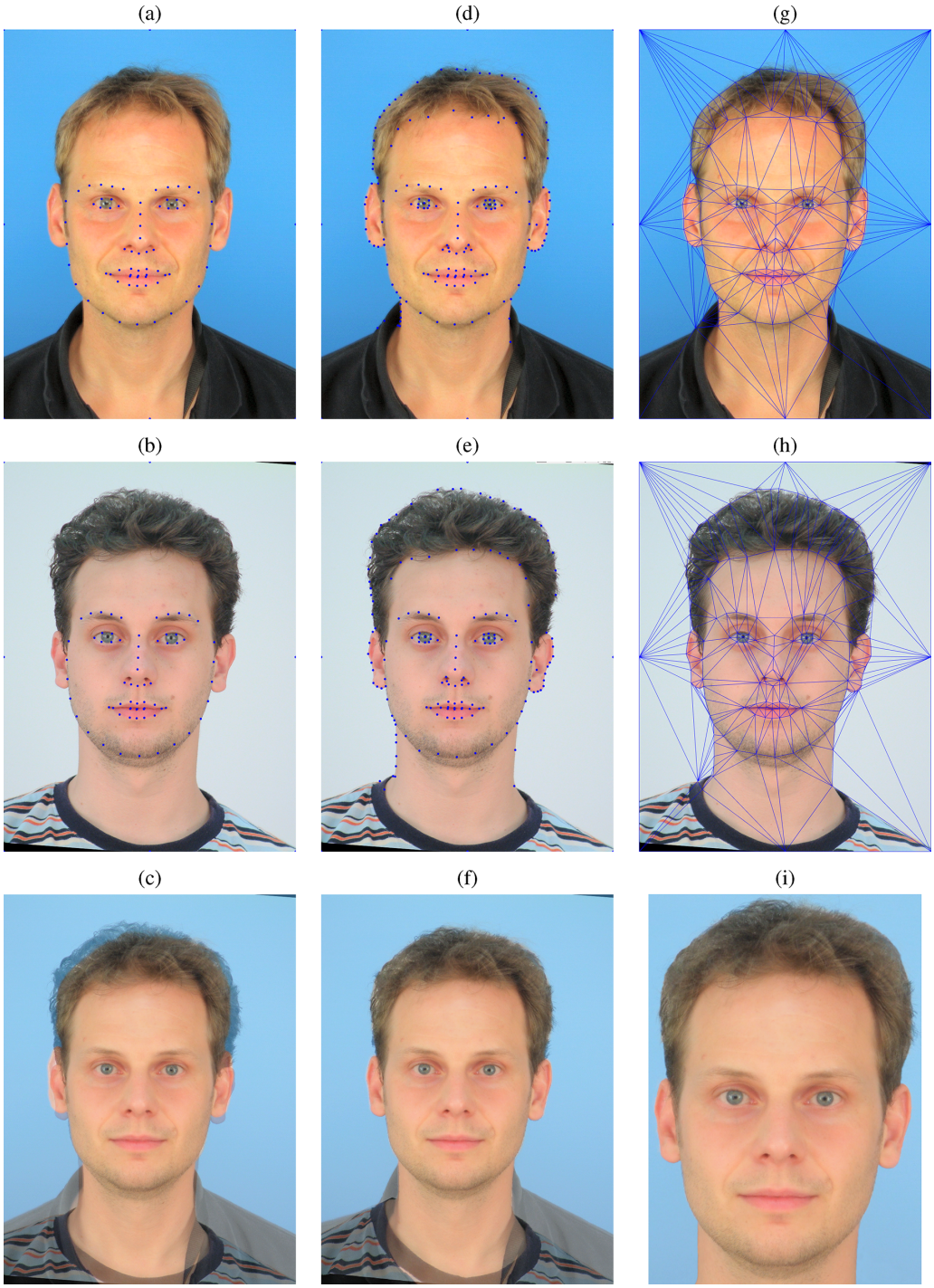

Figure 1. Shown are original faces of two different people f and g (a, b) with the automatically extracted facial landmark points overlaid (blue dots). Their midway morph m_{fg} generated using only these automatically extracted landmarks is shown in panel (c). Shown in panels (d) and (e) are the same original faces f and g, now with both the automatically and manually selected facial landmark points overlaid (blue dots). Shown in panels (g) and (h) are a tessellation of the faces used for the image morphing. The midway morph generated using both the automatically extracted and manually selected landmarks is shown in panel (f). Shown in panel (i) is the tightly cropped and manually touched-up and color and sharpness-adjusted final image. (Original image sources: Utrecht ECVP [Utrecht ECVP Face Database, 2008] and PUT face database [Kasiński et al., 2008].)

where the affine parameters a_1, a_2, a_3, and a_4 embody the anisotropic scaling, shearing, and rotation, and the parameters a_5 and a_6 embody the horizontal and vertical translation. The transform that best, in the least-squares sense, aligns the features on each face are estimated by first defining the following quadratic error function:

$$E(\vec{a}) = \left\| \begin{pmatrix} x_1 & y_1 & 0 & 0 & 1 & 0 \\ 0 & 0 & x_1 & y_1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{68} & y_{68} & 0 & 0 & 1 & 0 \\ 0 & 0 & x_{68} & y_{68} & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} - \begin{pmatrix} u_1 \\ v_1 \\ \vdots \\ u_{68} \\ v_{68} \end{pmatrix} \right\|^2 \quad (2)$$

$$= \|M\vec{a} - \vec{b}\|^2. \quad (3)$$

This quadratic error function is minimized using standard least-squares estimation: differentiate with respect to $\vec{a}$, set the result equal to 0, and solve for $\vec{a}$ to yield the least-squares estimate of the aligning affine transform:

$$\vec{a} = (M^t M)^{-1} M^t \vec{b}. \quad (4)$$

Given two aligned faces f and g (with VGG-representations $\vec{v}_f$ and $\vec{v}_g$), a morphed face m_{fg} is generated using a standard image-warping technique (Szeliski, 2010). Briefly, a triangular mesh is created on each face using the facial landmarks as vertices (Figure 1g, h). A midway morph is created by geometrically warping each triangular patch according to a morphing parameter $\alpha \in [0, 1]$, where a value of 0 corresponds to the source image f, a value of 1 corresponds to the source image g, and an intermediate value corresponds to a midway morph. The underlying pixel values are similarly computed as a weighted combination, $(1 - \alpha)f + \alpha g$, of the original pixel values (applied separately to each image color channel).

In past studies, morphs have often been generated with $\alpha = 0.5$, which means that the original faces f and g are weighted equally to generate a 50/50 morphed face m_{fg}. Previous research, however, has indicated that, when creating a morphed face, if one individual in the pair is more distinct than the other, then the 50/50 morph typically resembles the more distinct individual (Tanaka et al., 1998; Kramer et al., 2019). To compensate for this effect, we generated a range of morphs m^{α}_{fg} with α ranging from 0.1 to 0.9, in steps of 0.1. The value of α was selected that led to a morphed face m^{α}_{fg} that was, in the Euclidean sense on the underlying VGG-representation, midway between the source images f and g.

To improve overall contrast, each morph m^{α}_{fg} was gamma-corrected with $\gamma = 1.5$. The morphs were then tightly cropped around the face and manually edited to remove obvious morphing artifacts (Figure 1f vs. i). Lastly, to ensure that the morphing process did not create any obvious artifacts, the images were matched in terms of luminance, color, and sharpness. In particular, the mean luminance of each source image f and g was matched to the mean luminance of the morph image m_{fg}, and the source image mean and variance of the chrominance channels (Cb/Cr) were matched to the morphed image. Because image morphing tends to lead to blurring, each RGB color channel of each source and morph image was high-pass filtered until the average gradient of each image channel matched the maximum gradient across all three images. The resulting 54 pairs of different individuals and their midway morph comprise our *different-individual* data set.

An analogous *same-individual* data set was created by selecting a new set of 54 facial images from the original data set of 3,500 for which there were two or more distinct images of the same person. We manually selected individuals to match the gender, age, and race distribution of our different-individuals data set. A midway morph was created for each pair of images using the same technique described above.

In summary, our data set consists of 108 face pairs, 54 of two different individuals and 54 of the same individual taken at different times, each with a midway morph. A representative set of these different and same faces and morphs is shown in Figure 2.

Experiment 1a: Identification (original and morph)

In this first experiment, we examined people's ability to determine whether two facial images, one original and one morphed, are of the same person or not.[2]

Methods

One hundred workers on Amazon's Mechanical Turk (AMT) completed the experiment. The participants self-reported as 65 men, 34 women, and 1 prefer not to say; between 22 and 72 years of age ($\mu = 36.8$; $\sigma = 9.6$); and 74 White, 14 South Asian, 7 East Asian, and 5 African American. Participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20th percentile.

A within-subject design was employed in which each trial consisted of two images (one original, one morph) displayed side-by-side in one of eight configurations. Denote the original images of different people as f and g and their midway morph as m_{fg}, and denote the original images of the same people as h and $\tilde{h}$ and their midway morph as $m_{h\tilde{h}}$. There are four configurations for each data set consisting of 54 pairs from the "different-individual" data set with one image on the left and one on the right: $f + m_{fg}$, $m_{fg} + f$, $g + m_{fg}$,

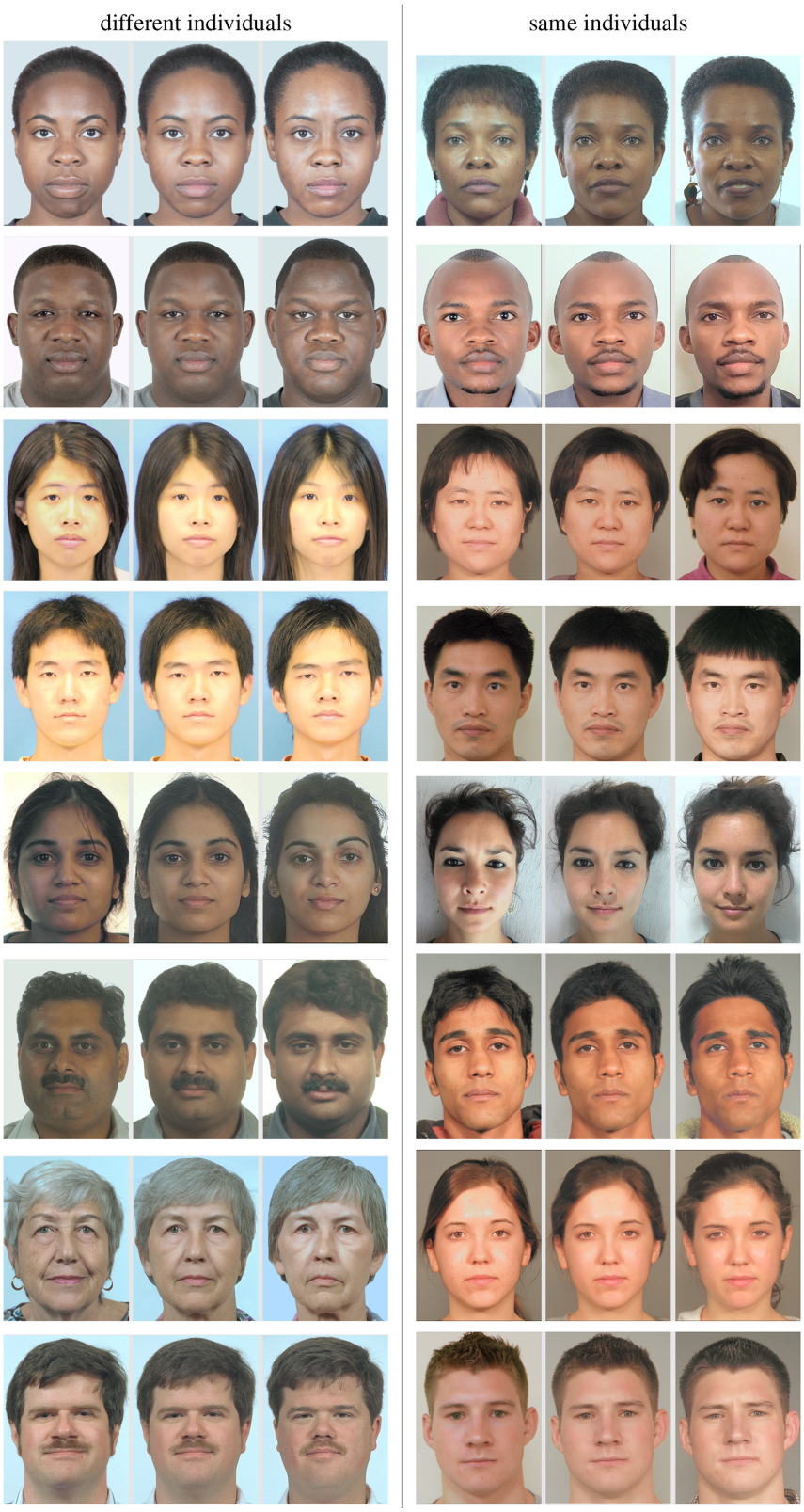

Figure 2. Example stimuli. Shown for each set of three images are two original images (left/right) and their midway morph (center). Shown in the left half of this figure are images from different-individual data set and shown in the right half are images from same-individual data set. (Original image sources: MR2 [CC BY-NC-SA 4.0], Chicago Face Database [permission to publish images granted], CUHK student database [Wang & Tang, 2008], NIST color FERET [image publication permitted under fair use policy], and CVRL ND-Collections B and D, and FRCG v.2.0 [image publication permitted under fair use policy].)

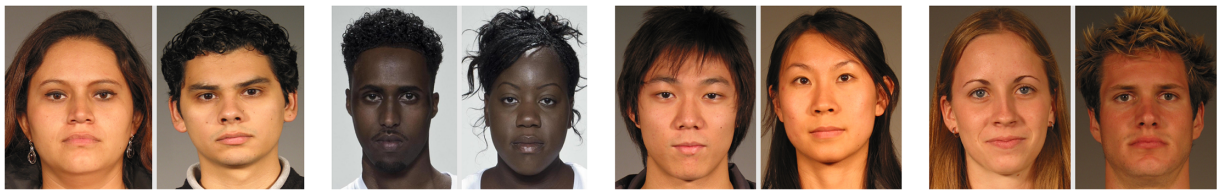

Figure 3. Catch trials used in Experiments 1a, 1b, and 1c to ensure that participants were paying attention to the task. (Original image sources: Face Research Lab London Set [CC BY 4.0] and CVRL ND-Collections B and D, and FRCG v.2.0 [image publication permitted under fair use policy].)

$m_{fg} + g$, and 54 pairs from the "same-individual" data set: $h + m_{h\tilde{h}}$, $m_{h\tilde{h}} + h$, $\tilde{h} + m_{h\tilde{h}}$, $m_{h\tilde{h}} + \tilde{h}$, for a total of 432 possible displays. Each participant viewed only 108 image pairs using the following fully counterbalanced block design. Four blocks were created, each containing 27 trials for a total of 108 trials. The first and second blocks each consisted of 14 different and 13 same image pairs; the third and fourth blocks each consisted of 13 different and 14 same image pairs. Each block consisted of the same number of men, women, and racial groups.

On each trial, participants were instructed to specify if the images were of the same person or not and asked to rate the confidence in their response. In this discrimination task, chance performance is 50%. Four attention-check trials were created, one for each block. These trials were intentionally easy, comprising two images of distinctly different-looking people, one male and one female (Figure 3).

Participants first received task instructions, including a brief description of what face morphing is and how it can be used to commit identity fraud. Participants then completed a practice trial; they viewed two images on the screen, an original image and a morph. Participants had an unlimited amount of time to indicate whether or not they thought that the images were of the same individual. After responding to the same/different individual question, participants rated their confidence in their decision using a 6-point Likert-type scale, from 1 (*50%—guessing*) to 6 (*100%—absolutely certain*).

Following the practice trial, participants completed the 108 trials in blocks of 27 plus one attention-check trial per block, shown in a randomized order within each block. Blocks were shown in one of four possible counterbalanced orders. At the end of the session, participants were asked a few basic demographic questions.

A precision-for-planning analysis revealed that at least 99 participants would provide a margin of error that is 0.2 of the population standard deviation with 95% assurance (Cumming et al., 2012; Cumming, 2013). This analysis applies to all reported experiments.

Results

The average accuracy of identifying a facial image as the same person or not was 59.2% (chance is 50%), corresponding to a sensitivity of $d' = 0.68$ and bias of $\beta = 1.81$, where the bias corresponds to a tendency to label faces as "same" (Table 1). The accuracy for faces of different/same individuals was 29.5%/88.8%—participants were heavily biased to saying that faces were of the same individual.

Experiment	d'	β	% correct [95% CIs]
1a	0.68	1.81	59.2 [57.6, 60.7]
1b	1.74	1.03	80.8 [78.8, 82.8]
1c	0.57	1.44	59.2 [57.9, 60.6]
2a	0.21	0.98	54.1 [52.5, 55.5]
2b	0.53	0.92	60.4 [58.9, 61.9]

Table 1. Participant accuracy in five experiments, reported as sensitivity (d'), bias (β), and accuracy (% correct with [95% confidence intervals]). Experiment 1a: determine if two images, one original and one morphed, are of the same person; Experiment 1b: determine if two images, both original, are of the same person; Experiment 1c: replication of Experiment 1a with training; Experiment 2a: determine if a single facial image is a morph or not; and Experiment 2b: replication of Experiment 2a with training and feedback.

Shown in Figure 4 (Experiment 1a) is, for each level of participant-reported confidence (on a scale of 1 to 6 [certain]), the participant accuracy. With similar average accuracy across all levels of confidence, we see that participants are not well calibrated in their response and confidence.

Discussion

The results of this experiment suggest that human participants have a limited ability to reliably determine the identity in a midway facial morph. There are two possible interpretations of this result: (1) The morphs are of high enough quality and similarity as to mask identity, or (2) participants are simply unable to accurately distinguish two unfamiliar faces. In the next experiment (1b), we seek to differentiate between these two possibilities by asking participants to distinguish between two original nonmorphed facial images consisting of the same or different person.

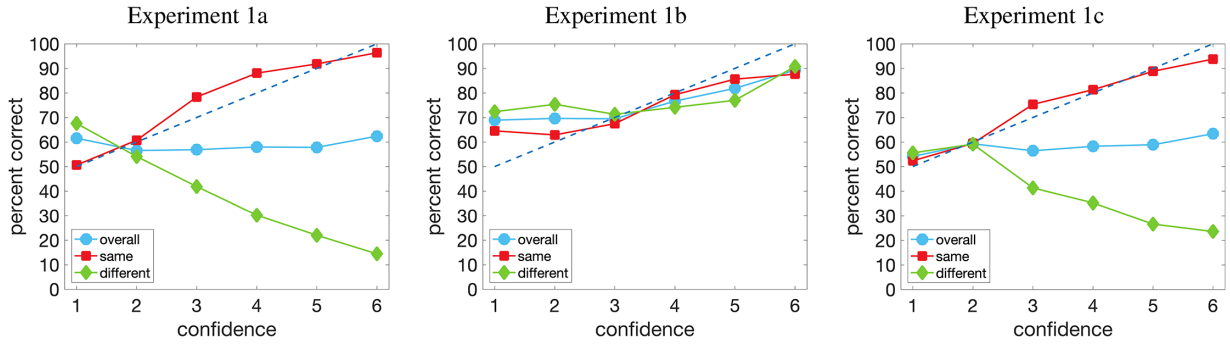

Figure 4. Confidence-accuracy curves for Experiments 1a, 1b, and 1c. The dashed line represents perfect accuracy-confidence calibration.

Experiment 1b: Identification (original and original)

Methods

One hundred workers on AMT completed the experiment. The participants self-reported as 53 men and 47 women; between 23 and 69 years of age ($\mu = 39.4$; $\sigma = 10.9$); and 74 White, 14 South Asian, 6 African American, 2 East Asian, and 4 other/prefer not to say. Participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20th percentile. A further two participants were excluded because they responded incorrectly on at least one of the attention-check questions. There was no overlap between the participants in this experiment and Experiment 1a.

A within-subject design was employed in which each trial consisted of two original images—from the same data set as used in Experiment 1a—displayed side-by-side in one of four configurations. Each participant saw 54 different individuals (f, g) with two possible configurations: $f + g$ or $g + f$, and 54 same individuals (h, $\tilde{h}$) with two possible configurations: $h + \tilde{h}$ or $\tilde{h} + h$, for a total of 216 possible displays. Each participant viewed 108 image pairs using the counterbalanced block design per Experiment 1a. Given that this experiment did not include any morphed facial images, we removed the brief explanation of face morphing from the task instructions and amended the practice trial to show two distinct original images of the same person. The procedure was otherwise identical to that of Experiment 1a.

Results

The average accuracy of identifying two facial images as depicting the same person or not was 80.8%, corresponding to a sensitivity of $d' = 1.74$ and bias of $\beta = 1.03$. The accuracy for faces of different/same individuals was 80.4%/81.3%—unlike the previous experiment, participants were not biased in their responses. Shown in Figure 4 (Experiment 1b) is, for each level of participant-reported confidence, the participant accuracy. With slightly higher accuracy at the higher levels of confidence, it appears that participants are fairly well calibrated in their response and confidence.

Discussion

The results of this experiment suggest that participants can reliably determine whether two original facial images depict the same person or two different people. Therefore, participants' limited ability in the task in Experiment 1a may be interpreted as a result of the morphs being of high enough quality and similarity to mask identity.

Given that participants can distinguish two unfamiliar faces with reasonable accuracy, we next examine whether participant accuracy in distinguishing identity in morphed faces can be improved with training. To develop our training initiative, we draw on the finding that attending to certain facial features when comparing two faces can help to improve the accuracy of face matching decisions (Kemp et al., 2016; Towler et al., 2017). In the next experiment (1c), we replicate Experiment 1a but this time using masked facial images that allow participants to only compare the eyes, nose, and mouth.

Experiment 1c: Identification (original and morph) with masking

Methods

One hundred workers on AMT completed the experiment. The participants self-reported as 51 women,

Figure 5. Example masked stimuli from Experiment 1c. Shown in each image pair is an original image of an individual f (left) and the midway morph m_{fg} (right) to a different person g, with only the eye region or mouth and nose region visible. (Original image source: Chicago Face Database [permission to publish images granted].)

48 men, and 1 prefer not to say; between 22 and 65 years of age ($\mu = 40.3$; $\sigma = 9.6$); and 81 White, 7 East Asian, 6 African American, 5 South Asian, and 1 other/prefer not to say. As in the previous two experiments, participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20th percentile. There was no overlap between the participants in this experiment and Experiments 1a and 1b.

Other than the two exceptions enumerated below, the design, underlying stimuli, and procedure were identical to that used in Experiment 1a.

(1) In each trial, participants saw a pair of images (one original and one morph) displayed side-by side. One image pair revealed only the eyes, and the other image pair revealed only the nose/mouth region (Figure 5). The creation of these masks was automated as follows. The pixel locations of the facial features (eyes, nose, mouth) were extracted using OpenFace (Baltrušaitis et al., 2016). For the eyes, a bounding box was extracted that contained all of the features on both eyes. For the nose/mouth, a bounding polygon was extracted that contained all of the features on the nose and mouth. To ensure that the mask did not occlude the features, these bounding boxes were enlarged by 5% of their original size. The final images (Figure 5) were generated by reducing the contrast of all pixels outside of the mask to 15% of full contrast, making it difficult for participants to use the entire face for recognition, while leaving enough context for the visible features.

(2) The order in which participants viewed the two feature regions was counterbalanced, resulting in twice the number of display configurations as in Experiment 1a. Participants saw 27 of each of the different people configurations ($f + m_{fg}$, $m_{fg} + f$, $g + m_{fg}$, $m_{fg} + g$) with the eye region shown first and the mouth and nose region shown second and 27 with the mouth and nose region shown first and the eye region shown second. Participants saw 27 of each of the same people configurations ($h + m_{h\tilde{h}}$, $m_{h\tilde{h}} + h$, $\tilde{h} + m_{h\tilde{h}}$, $m_{h\tilde{h}} + \tilde{h}$) with the eye region shown first and the mouth and nose region shown second and 27 with the mouth and nose region shown first and the eye region shown second. Each participant viewed a total of 108 image pairs.

Results

The average accuracy of identifying a facial image as the same person or not was 59.2%, corresponding to a sensitivity of $d' = 0.57$ and bias of $\beta = 1.44$ (cf. Experiment 1a: 59.2%, $d' = 0.68$, $\beta = 1.81$; Table 1). The accuracy for faces of different/same individuals was 36.1%/82.3%—although overall accuracy was still not high, the masking reduced the bias to report that faces were of the same individual. As in Experiment 1a, average participant accuracy is similar across all levels of confidence, suggesting that participants are still not well calibrated in their response and confidence (Figure 4 [Experiment 1c]).

Discussion

Had the strategy of focusing participants' attention on facial features been successful in increasing accuracy or decreasing bias, this would have been a simple strategy for a passport issuance office to adopt. The results of this experiment, however, reveal that facial-feature comparison did not significantly improve participants' accuracy in determining identity of morphed faces. Compared to Experiment 1a, however, participants showed a smaller bias to respond "same." Participants clearly struggle to distinguish identity

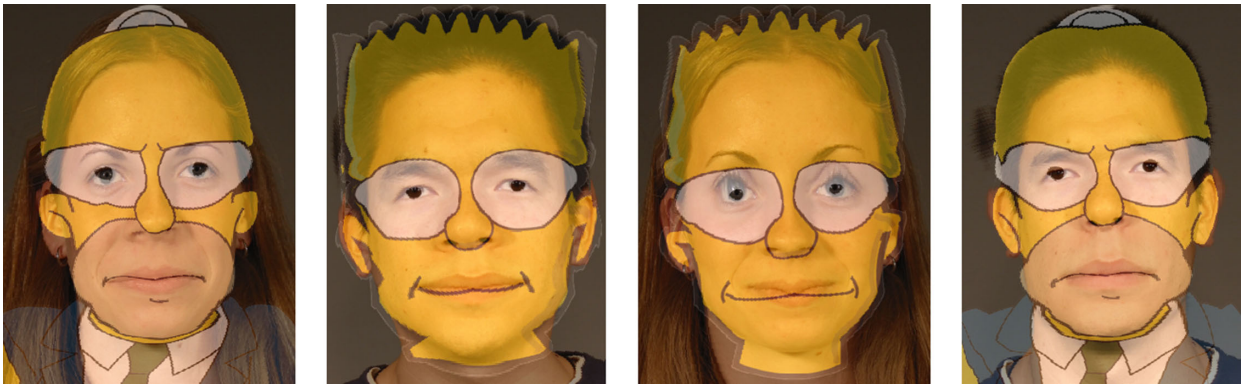

Figure 6. Catch trials used in Experiments 2a and 2b to ensure that participants were paying attention to the task. (Original image sources: CVRL ND-Collections B and D, and FRCG v.2.0 [image publication permitted under fair use policy] and Pixy.org [CC BY-NC-ND 4.0].)

when presented with two images, one of which is a midway morph.

Distinguishing identity, however, is only one way in which a fraudulent identity might be determined. The other way is to simply identify a face as having been morphed relative to some unknown face. In the next set of experiments, we examine participants' ability to perform this task.

Experiment 2a: Classification (original or morph)

Methods

One hundred workers on AMT completed the experiment. The participants self-reported as 57 men and 43 women; between 24 and 72 years of age ($\mu = 40.4$; $\sigma = 10.2$); and 78 White, 10 South Asian, 5 African American, 4 East Asian, and 3 other/prefer not to say. Participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20th percentile. A further three participants were excluded because they responded incorrectly on at least one of the attention-check questions. There was no overlap between the participants in this experiment and Experiments 1a, 1b, or 1c.

A within-subject design was employed in which each trial consisted of a single original or morphed face. For the morphed face trials, each participant saw 27 different-people midway morphs (m_{fg}) and 27 same-person midway morphs ($m_{h\tilde{h}}$). For the original image trials, participants saw 27 images, either f or g, from the different-individual image pairs and 27 images, either h or $\tilde{h}$, from the same-individual image pairs.

Each participant viewed 108 images using the following fully counterbalanced block design. Four blocks were created, each containing 27 trials for a total of 108 trials. The first and second blocks each consisted of 14 original face trials and 13 morphed face trials; the third and fourth blocks each consisted of 14 morphed face trials and 13 original face trials. The selection of the original or midway morph from each of the 108 image sets was also counterbalanced, resulting in two versions of each block.

On each trial, participants were instructed to specify if the image was a morph or not and asked to rate the confidence in their response. In this task, chance performance is 50%. Four attention-check trials were created, one for each block. These trials were intentionally easy, comprising a morphed face of a person with an image of a cartoon character (Figure 6).

Participants first received task instructions, including a brief description of what face morphing is and how it can be used to commit identity fraud. Participants then viewed four videos demonstrating how two faces can be digitally combined to create a morph of those two faces. To create these videos, we selected four additional face pairs from the original data set of 3,500 faces. For each face pair, we generated morphs using the same method as described previously (see Data set section). To demonstrate the gradual morphing of two faces, we generated five morphs with a different blending value α ranging from 0.1 to 0.5 in steps of 0.1. The $0.5 - \alpha$ morph was then manually edited to remove obvious morphing artifacts. In the video, the two original images (f and g) were displayed on either side of their morph (m_{fg}). The six versions of the morph appeared sequentially, starting with the $0.1 - \alpha$ morph. Each version of the morph remained on the screen for 1 s. Participants viewed all four videos and were asked to indicate if they were able to see the videos clearly.

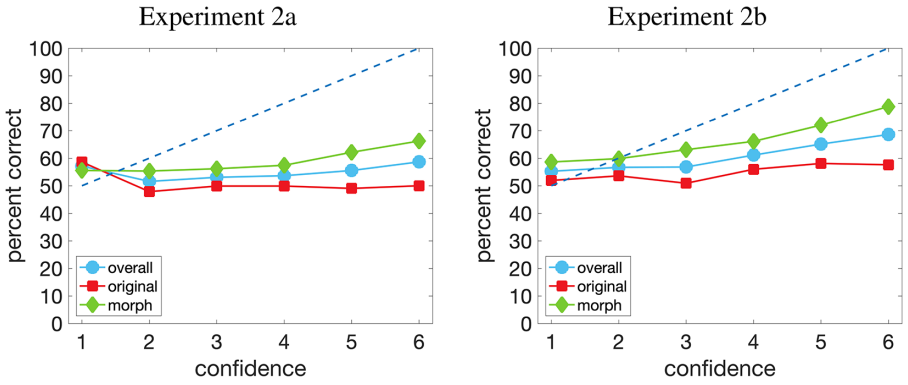

Figure 7. Confidence-accuracy curves for Experiments 2a and 2b. The dashed line represents perfect accuracy-confidence calibration.

Participants then completed a practice trial consisting of a single image on the screen, an original image, or a midway morph.

Following the practice trial, participants completed the 108 trials in blocks of 27 plus one attention-check trial per block, shown in a randomized order within each block. Blocks were shown in one of eight possible counterbalanced orders. At the end of the session, participants were asked a few basic demographic questions.

Participants had an unlimited amount of time to indicate whether they thought that the image was a morph or not. After responding to the morph or not question, participants rated their confidence in their decision using a 6-point Likert-type scale, from 1 (*guessing*) to 6 (*absolutely certain*).

Results

The average accuracy of identifying a face as a morph or not was 54.1%, corresponding to a sensitivity of $d' = 0.21$ and bias of $\beta = 0.98$. The accuracy for original/morphed faces was 50.0%/58.1%. As in Experiments 1a and 1c, average participant accuracy was similar across all levels of confidence, again suggesting that participants are not well calibrated in their response and confidence (Figure 7 [Experiment 2a]).

Discussion

The results of this experiment suggest that human participants cannot reliably determine when a facial image has been morphed and when it is an original image. Two possible explanations for this result are that (1) the morphed faces are of high enough quality that there are no artifacts that can be reliably perceived by human participants, or (2) participants are unaware of the artifacts to look for in morphed faces. To try to determine which of these two possibilities best accounts for our results, in the next experiment, we replicate Experiment 2a, but before attempting the task, participants completed a short training session that highlights some common morphing artifacts to look for in the images.

Experiment 2b: Classification (original or morph) with training/feedback

Methods

One hundred workers on AMT completed the experiment. The participants self-reported as 58 men, 41 women, and 1 prefer not to say; between 24 and 68 years of age ($\mu = 39.7$; $\sigma = 9.9$); and 81 White, 9 South Asian, 4 African American, 2 East Asian, and 4 other/prefer not to say. As in the previous experiments, participants received $5 for completing this experiment. As an incentive to encourage effort on the task, a $5 bonus was offered and paid for those achieving an accuracy in the top 20th percentile. A further five participants were excluded because they responded incorrectly on at least one of the attention-check questions. There was no overlap between the participants in this experiment and Experiments 1a, 1b, 1c, and 2a.

Other than the inclusion of a training session as described below and accuracy feedback after each trial, the design, underlying stimuli, and procedure were identical to that used in Experiment 2a.

After viewing the four videos demonstrating how two faces can be digitally combined to create a morph of those two faces, participants completed a short training. The training provided information about common morphing artifacts that may be helpful to look for when deciding if the facial images were morphs or not. Participants were told that:

(1) Morphed faces tend to look less sharp: The complexion of a morphed face is usually smoother with a more uniform appearance.

(2) Morphing hair can be difficult, and morphs often have fewer wayward strands of hair and ghosting (a ghost-like outline of another person's hair).
(3) When morphing images of two people with different postures or different clothing/hair coverage, the editing process might make the neckline appear unnaturally straight and flat.

To check that participants paid attention to the training, they were then asked to select the three artifacts from a list of six possible options, where the three incorrect answers were easily identifiable as they were not mentioned in the training. Participants were given the option to view the training session a second time if they were unsure of the correct options. The other change from Experiment 2a was that after responding on each trial, participants were provided with feedback indicating whether their response was correct or not.

Results

The average accuracy of identifying a face as a morph or not was 60.4%, corresponding to a sensitivity of $d' = 0.53$ and bias of $\beta = 0.92$ (cf. Experiment 2a: 54.1%, $d' = 0.21$, $\beta = 0.98$; Table 1). The accuracy for original/morphed faces was 54.6%/66.2%. Average accuracy was only slightly higher at the higher levels of confidence (4–6) than the lower levels (1–3) of confidence, suggesting participants had a limited ability to calibrate their response and confidence (Figure 7 [Experiment 2b]).

Discussion

The results of this experiment indicate that raising awareness of morphing artifacts and providing feedback led to only a small improvement in participants' accuracy in determining whether a facial image has been morphed or not. Even with this training, participants struggled to reliably identify a face as having been morphed relative to an unknown face.

Taken together, the results of our five experiments suggest that people are unable to reliably detect face morphing, neither by distinguishing identity nor by classifying a face as having been morphed. Next we examine whether computational approaches can be used to detect face morphing.

Discussion: Experiments 1 and 2

Crowd wisdom

To determine whether groups are more accurate in the detection of face morphing than individual decision makers, we next examined whether there is "wisdom in the crowd" (Hastie & Kameda, 2005). For each experiment, we aggregated the 100 participant responses for each of the 108 trials using a majority rules criterion.

Using this crowd-based approach in Experiment 1a resulted in an average accuracy of 53.8% for identifying a facial image as the same person or not, which was not reliably different from the average of individual responses (59.2%). In Experiment 1b, however, average accuracy using the crowd-based approach was 16.4% higher than the averaged individual responses (97.2% vs. 80.8%, 95% CI [12.7%, 20.1%]). In addition, in Experiment 1c, where participants received training, the crowd-based approach resulted in an average accuracy similar to the individual approach (58.3% vs. 59.2%). The improved accuracy in Experiment 1b suggests that participants make different mistakes, and so pooling across multiple responses improves overall accuracy.

In Experiment 2a, accuracy in classifying images as original or morphs was similar when averaging individual (54.1%) and crowd (58.4%) responses. When participants received training and feedback (Experiment 2b), the crowd-based approach resulted in an average accuracy 12.2% higher than the individual approach (60.4% vs. 72.6%, 95% CI [6.1%, 18.4%]). Without any training (Experiment 2a), participants typically made the same mistakes, but having received training (Experiment 2b), there was greater variation in which of the trials participants responded correctly on. This result suggests that with training and feedback, the crowd becomes wiser.

It is possible that aggregating across identity verification decisions of multiple passport officers might lead to greater accuracy in real-world passport issuance. Of course, this additional effort might not be feasible, and we also note that when people know a decision is group based, it can lead to social loafing (Hastie & Kameda, 2005).

Computational identification

The results of Experiments 1a and 1c show that participants' ability to determine whether two facial images, one original and one morphed, are of the same person or not is limited. We next examine whether computational techniques can perform this task. A standard convolutional neural network (Parkhi et al., 2015) was used to extract a low-dimensional, perceptually meaningful (Tariq et al., 2018; Zhang et al., 2018) representation of each face in our data set of 108 face pairs and their corresponding midway morph. This is the same VGG representation used earlier to determine the similarity between two faces and to compute the midway morph.

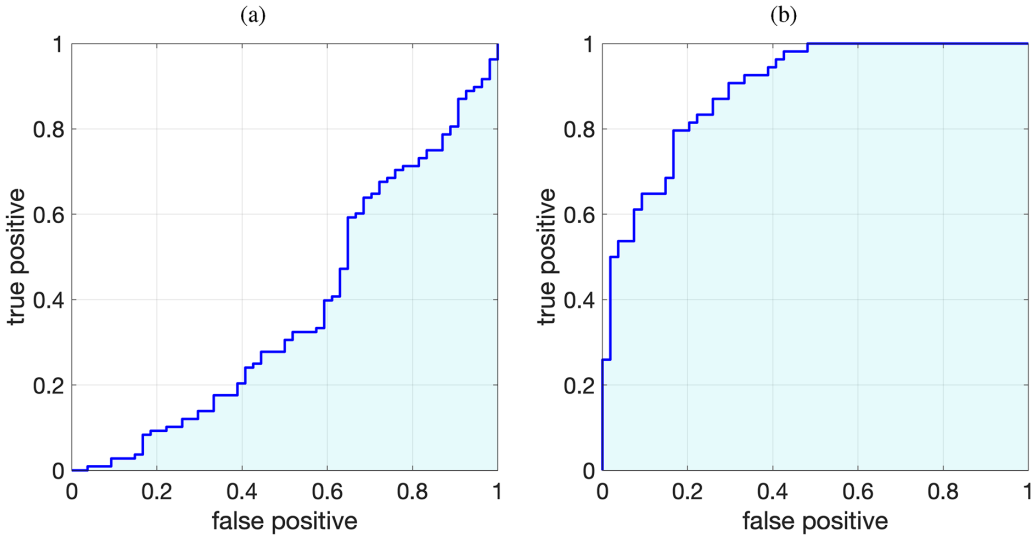

Figure 8. Receiver-operating-characteristic curves for (a) same individuals (true positive) and different individuals and their midway morph (false positive). (b) Same individuals (true positive) and different individuals (false positive).

For each of the 54 pairs of faces of the same individual taken at different times (h and $\tilde{h}$), the similarity between these faces was measured as the Euclidean distance between the VGG representation of the two original faces. Similarly, the distance was computed between the VGG representation for each of the 54 pairs of different individuals (f and g) and their midway morph (m_{fg}).

Shown in Figure 8(a) is the receiver operating characteristic (ROC) curve plotted as the true-positive rate (correctly identifying the same individual) as a function of the false-positive rate (incorrectly identifying an individual and their midway morph as the same). The area under the curve (AUC) is 0.38, where a chance classifier would have an AUC of 0.5, showing that even a state-of-the-art, machine-learning, face recognition algorithm is not able to perform this identification task. We note that, in a flipped classifier, this result effectively corresponds to an AUC of 0.62, still illustrating a fairly limited performance. Interestingly, however, the below-chance AUC result indicates that a morphed face is highly similar to the source faces, a finding that we draw on in the subsequent Computational Classification section.

We next evaluate if this computational approach can perform the task of face recognition outside of the issue of morphing (as in Experiment 1b). The distance was computed between the VGG representation for each of the 54 pairs of faces of two different individuals (f and g). Shown in Figure 8b is the ROC for this task, now with an AUC of 0.90. Although VGG-based facial recognition is generally effective in distinguishing between different individuals, it struggles to distinguish between morphed faces, reinforcing just how difficult this task is.

Facial recognition systems have been shown to perform worse at recognizing faces of women and Black individuals (Klare et al., 2012; Buolamwini & Gebru, 2018). We next examined whether these biases were present when using the VGG-based face recognition algorithm to perform our identification task. When identifying pairs of faces of the same individual taken at different times (h and $\tilde{h}$) and the 54 pairs of different individuals (f and g) and their midway morph (m_{fg}), the face recognition algorithm performed worse on Black faces (AUC = 0.00) than East Asian (AUC = 0.29), South Asian (AUC = 0.61), or White (AUC = 0.29) faces. The algorithm performed slightly better for women (AUC = 0.43) than for men (AUC = 0.33). In addition, in the absence of morph faces, the VGG-based facial recognition performed worse for Black faces (AUC = 0.75) than for the other races (all AUCs > 0.90). There was no difference in face recognition performance for women and men.

Computational classification

In the previous sections, we saw that a midway morph between two different people looks similar enough to each person so as to cause consistent misidentification by human participants and state-of-the-art facial recognition. In this section, we attempt to leverage the unusual similarity between two photos of a person as a possible indication that one of the photos is a midway morph.

Consider, for example, the use of a midway morph in identity theft in which person f is attempting to steal person g's identity and submits a request for a

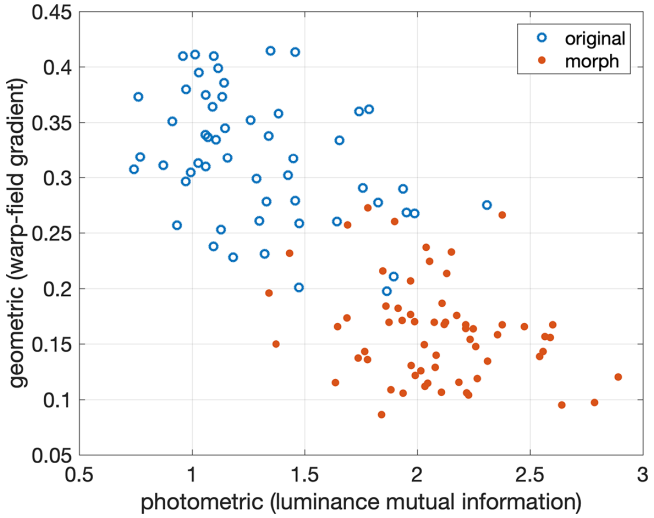

Figure 9. Photometric and geometric measurements between two distinct images of the same individual (original) and an image of an individual and a midway morph to another individual (morph). A lower geometric measure corresponds to a higher degree of similarity, and a higher photometric measure corresponds to a higher degree of similarity.

new passport with a midway morph photo m_{fg}. The passport office will compare the original photo g with the new photo m_{fg} to make sure that it is the same person. Per our earlier results, and assuming a high-quality morph, the faces will look similar enough to match. The two photos g and m_{fg}, however, will share significant geometric and photometric properties because the morphed image m_{fg} is composed of one half of the original image g, as compared to a completely new photo of person g, which will almost certainly differ somewhat in terms of head pose, facial expression, lighting, and so on.

We hypothesize, therefore, that a pair of images of an individual, one of which is a morph, will be more geometrically and photometrically similar than two separately photographed images of an individual. Each of the original images was registered to the morphed image using a standard local and nonrigid registration (using MATLAB's `imregdemons`), parameterized as a local two-dimensional motion field (v_x, v_y). The magnitude of the geometric distortion between two images is quantified as the average magnitude of the gradient of the underlying motion field $(\sqrt{v_x^2 + v_y^2})$. Once aligned, the photometric similarity between two images is quantified as the mutual information (Cover & Thomas, 2012) on the luminance channel.

Shown in Figure 9 are the geometric and photometric measurements for morphed (filled red circles) and original (open blue circles) images. Each original data point corresponds to the average difference h aligned to $\tilde{h}$, and $\tilde{h}$ aligned to h, where h and $\tilde{h}$ correspond to distinct images of the same person. Each morphed data point corresponds to the average difference between f aligned to m_{fg} and g aligned to m_{fg}, where f and g correspond to images of different people. The morphed images are distinctly more photometrically similar (having a higher luminance mutual information) and more geometrically similar (having a smaller warp-field gradient). This, again, is as it should be given how the morphed image is created.

Because all of the images in our data set are passport-style photos, this unusually high similarity among the morphed images is not simply an artifact of the style of the photographs. This unusual similarity can, therefore, be used as a cue to flag potentially suspiciously similar images.

General discussion

We have examined both perceptual and computational approaches for detecting face morphing. Across five perceptual studies, human participants showed a limited ability to detect face morphing, both by distinguishing identity (Experiment 1a) and by classifying a morphed face (Experiment 2a). Training did not significantly improve performance in the identification task (Experiment 1c), and training and feedback resulted in only a small improvement in performance in the classification task (Experiment 2b). Additionally, we found that even a state-of-the-art, machine-learning, face recognition algorithm could not reliably distinguish one person from a midway morph. We did, however, identify a computational technique to leverage the unusual similarity between a pair of images when one is a midway morph. This technique could be implemented at passport issuance to help in flagging suspicious applications for further processing.

Our results are in line with previous work showing that human ability to detect face morphing is limited (Robertson et al., 2017; Robertson et al., 2018; Kramer et al., 2019). Our results, showing that training participants to detect morphing artifacts has a limited effect on performance when using high-quality face morphs, are concordant with the findings of Kramer et al. (2019). Although facial masking has been found to improve performance in unfamiliar face matching tasks (Kemp et al., 2016), we note that this strategy did not have a reliable effect on participants' ability to distinguish identity in morphed faces; it did, however, reduce participants' bias to respond that two faces were of the same individual. That said, our approach involved masking the face to only reveal certain facial features (eyes, nose, and mouth) and does not provide a direct replication of the method used in previous research where the outer contour of the face was hidden (Kemp et al., 2016). Future investigations

might examine the effectiveness of a training approach that, rather than masking the faces, involves either removing the outer contour (Kemp et al., 2016) or asking participants to rate the similarity of certain facial features (Towler et al., 2017).

In the absence of morphing (Experiment 1b), there is wisdom in the crowd not seen in individual responses. One possible reason for this crowd wisdom is that there are both individual and intraindividual differences in face-processing strategies and abilities (Hong & Page, 2004; Webster et al., 2004; Bindemann et al., 2012; White et al., 2013). According to the diagnostic feature-detection hypothesis, people's ability to accurately distinguish between two faces varies according to the extent to which they rely on diagnostic features (features that differ across faces) or nondiagnostic features (features that appear the same across faces) (Gibson, 1969; Mundy et al., 2007; Wixted & Mickes, 2014). Even when participants consider and compare specific features, the individual reliance placed on the most diagnostic features during the task varies across individuals, thereby creating differences in per trial accuracy. Moreover, individual decision-making is affected by context (Çelen et al., 2004; Kremer et al., 2014). In our case, participants' sequential decision-making might be influenced by factors carrying over from the previous trial. In other words, it is possible that the ability to use diagnostic facial features (signal) and ignore the similar features (noise) varies not only across individuals but also that there is variation in how an individual applies this strategy over time. It is likely that these differences in decision-making are also influenced by contextual factors, including an effect of feedback (Kremer et al., 2014), which might help to explain why there is wisdom in the crowd in Experiment 2b but not in 2a. It will be interesting to explore morph recognition and detection by experts in face perception (e.g., trained forensic examiners) and superrecognizers, who perhaps rely more on diagnostic than nondiagnostic facial features (Carey, 1992; Furl et al., 2002; Hills & Lewis, 2011; Kemp et al., 2016; Towler et al., 2017; Phillips, 2018).

In the real world, fraudsters attempting to use morphed facial images to commit identity fraud are likely to try to find the most similar-looking accomplice and take the time to generate a high-quality morph with minimal visible artifacts. It seems, therefore, that the possible advantage of training passport officers to look for certain artifacts may be of limited value. As technological advances allow for increasingly more sophisticated face morphs, it seems reasonable to not rely entirely on human review and move toward other interventions.

Given the difficulty of detecting face morphing paired with the threat posed by not detecting this type of fraud, it is not surprising that calls have been made to modify the passport issuance process. Specifically, some researchers have suggested that the best solution to the face-morphing problem is to have government officials acquire photos at the place of issuance (Ferrara et al., 2014). In fact, reports suggest that Germany might already be preparing to take such a move (Huggler, 2020). Although this would solve the problem of digital face morphing, it is likely to be a costly change and still does not deal with the issue of physical identity fraud techniques, such as the use of hyperrealistic silicon masks (Sanders, 2017).

Concluding on a more hopeful note, in the current article, we have identified one limitation of the morphing process that can be leveraged as a way to flag suspicious images at the issuance stage using a computational classification technique. The proposed technique is based on the assumption that the morphed facial image submitted with an application for a new identity document has been generated using the original facial image that the issuance office has on record with some unknown face. Using the image that is stored on record results in the newly submitted face morph image being too similar—photometrically and geometrically—to the image on record than would be found in photos of the same individual taken at two different times. This technique is low cost and easy to implement for flagging suspicious applications for further processing. Given the high error rates of human participants and the limited effect of training, this and other computational approaches might provide a more reliable and practical method for detecting face morphing.

Keywords: facial morphing, morph detection, identity fraud

Acknowledgments

Funding from Facebook, Google, and the Defense Advanced Research Projects Agency (DARPA FA8750-16-C-0166). The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. government.

Commercial relationships: none.
Corresponding author: Sophie J. Nightingale.
Email: s.nightingale1@lancaster.ac.uk.
Address: Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK.

Footnotes

[1]All images and morphs will be made available upon request.
[2]All experiments reported in this article were approved by University of California Berkeley's Office for Protection of Human Subjects (OPHS), Protocol ID: 2019-07-12422. Participants gave fully informed consent prior to taking part.

References

Bastanfard, A., Nik, M. A., & Dehshibi, M. M. (2007). Iranian Face Database with age, pose and expression. *Proceedings of the IEEE International Conference on Machine Vision (ICMV 2007),* 50–55, doi:10.1109/ICMV.2007.4469272.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. *Proceedings of the IEEE Conference on Applications of Computer Vision (WACV),* 1–10, doi:10.1109/WACV.2016.7477553.

Bindemann, M., Avetisyan, M., & Rakow, T. (2012). Who can recognize unfamiliar faces? Individual differences and observer consistency in person identification. *Journal of Experimental Psychology: Applied, 18*(3), 277.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification *Proceedings of the Machine Learning Research (MLR) 1st Conference on Fairness, Accountability and Transparency, 81*, 77–91.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286–291.

Carey, S. (1992). Becoming a face expert. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 335*(1273), 95–103.

Çelen, B., & Kariv, S. (2004). Distinguishing informational cascades from herd behavior in the laboratory. *American Economic Review, 94*(3), 484–498.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). John Wiley & Sons.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*(3), 138–146.

DeBruine, L., & Jones, B. (2017). Face Research Lab London Set. Retrieved from https://figshare.com/articles/Face_Research_Lab_London_Set/5047666, doi:10.6084/m9.figshare.5047666.v3.

Ferrara, M., Franco, A., & Maltoni, D. (2014). The magic passport. *Proceedings of the IEEE International Joint Conference on Biometrics,* 1–7, doi:10.1109/BTAS.2014.6996240.

Flynn, P. J., Bowyer, K. W., & Phillips, P. J. (2003). Assessment of time dependency in face recognition: An initial study. In J. Kittler, & M. S. Nixon (Eds.), *Lecture Notes in Computer Science: Vol. 2688. Audio- and Video-Based Biometric Person Authentication (AVBPA)* (pp. 44–51). Springer, doi:10.1007/3-540-44887-X_6.

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect: Computational mechanisms for a developmental contact hypothesis. *Cognitive Science, 26*(6), 797–815.

Gibson, E. J. (1969). Principles of perceptual learning and development. Appleton-Century-Crofts.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review, 112*(2), 494.

Hills, P. J., & Lewis, M. B. (2011). Reducing the own-race bias in face recognition by attentional shift using fixation crosses preceding the lower half of a face. *Visual Cognition, 19*(3), 313–339.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences, 101*(46), 16385–16389.

Huggler, J. (2020). German photography studios protest government's new planned passport rules. Retrieved from https://www.telegraph.co.uk/news/2020/01/08/german-photography-studios-protest-governments-new-planned-passport/.

ITW Security Division. (2017). Fraudulently obtained genuine (FOG) documents: An ITW Security Division White Paper. Retrieved April 16, 2020, from https://www.itwsecuritydivision.com/Portals/0/documents/ITW%20White%20Paper%20-%20Fraudulently%20Obtain%20Genuine%20FOG%20Documents.pdf.

Kasiński, A., Florek, A., & Schmidt, A. (2008). The PUT Face Database. *Image Processing and Communications*, *13*, 59–64.

Kemp, R. I., Caon, A., Howard, M., & Brooks, K. R. (2016). Improving unfamiliar face matching by masking the external facial features. *Applied Cognitive Psychology, 30*(4), 622–627.

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research, 10*, 1755–1758.

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

Kramer, R. S., Mireku, M. O., Flack, T. R., & Ritchie, K. L. (2019). Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications, 4*(1), 28.

Kremer, I., Mansour, Y., & Perry, M. (2014). Implementing the "wisdom of the crowd." *Journal of Political Economy, 122*(5), 988–1012.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122–1135.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175–1184.

Middleton, R. (2014). For terrorists, documents are as important as weapons. *CSEye: Journal of the UK Forensic Science Society, 1*(2), 6–10.

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*(2), 124.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In X. Xie, M. W. Jones, & G. K. L. Tam (Eds.), *Proceedings of the British Machine Vision Conference (BMVC)* (pp. 41.1–41.12). BMVA Press, doi:10.5244/C.29.41.

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

Phillips, J. P., 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, J. P., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing, 16*(5), 295–306.

Phillips, J. P., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., 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.

Robertson, D. J., Kramer, R. S., & Burton, A. M. (2017). Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PLoS One, 12*(3), e0173319.

Robertson, D. J., Mungall, A., Watson, D. G., Wade, K. A., Nightingale, S. J., & Butler, S. (2018). Detecting morphed passport photos: A training and individual differences approach. *Cognitive Research: Principles and Implications, 3*(1), 1–11.

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: Principles and Implications, 2*(1), 1–12.

Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., & Heagins, T. B. (2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. *Behavior Research Methods, 48*(3), 1197–1204.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer, doi:10.1007/978-1-84882-935-0.

Tanaka, J., Giles, M., Kremen, S., & Simon, V. (1998). Mapping attractor fields in face space: The atypicality bias in face recognition. *Cognition, 68*(3), 199–220.

Tariq, T., Tursun, O. T., Kim, M., & Didyk, P. (2018). Why are deep representations good perceptual quality features? arXiv: 1812.00412.

Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing, 28*(6), 902–913.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*(1), 47.

UK HM Passport Office. (2020). Basic passport checks. Retrieved April 16, 2020, from https://www.gov.uk/government/publications/basic-passport-checks.

Utrecht ECVP Face Database. (2008). Retrieved April 7, 2020, from http://pics.stir.ac.uk/2D_face_sets.htm.

Vieira, T. F., Bottino, A., Laurentini, A., & De Simone, M. (2014). Detecting siblings in image pairs. *The Visual Computer, 30*(12), 1333–1345, doi:10.1007/s00371-013-0884-3.

Wang, X., & Tang, X. (2008). Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(11), 1955–1967.

Watson, C. I. (2010). *Multiple Encounter Dataset I (MEDS-I)*. (NISTIR, 7679). [Data set]. NIST.

Retrieved from https://www.nist.gov/publications/multiple-encounter-dataset-i-meds-i.

Webster, M. A., Kaping, D., Mizokami, Y., & Duhamel, P. (2004). Adaptation to natural facial categories. *Nature, 428*(6982), 557–561.

Weyrauch, B., Heisele, B., Huang, J., & Blanz, V. (2004). Component-based face recognition with 3D morphable models. *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, doi:10.1109/CVPR.2004.315.

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. *Applied Cognitive Psychology, 27*(6), 769–777.

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS One 9*. 1–6, doi:10.1371/journal.pone.0103510.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 586–595, doi:10.1109/CVPR.2018.00068.