Published in final edited form as:

ACM Trans Appl Percept. 2024 January ; 21(1): . doi:10.1145/3618113.

The Influence of the Other-Race Effect on Susceptibility to Face Morphing Attacks

SNIPTA MALLICK,
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX

GÉRALDINE JECKELN,
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX

CONNOR J. PARDE,
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX

CARLOS D. CASTILLO,
Whiting School of Engineering, Johns Hopkins University, College Park, MD

ALICE J. O'TOOLE
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX

Abstract

Facial morphs created between two identities resemble both of the faces used to create the morph. Consequently, humans and machines are prone to mistake morphs made from two identities for either of the faces used to create the morph. This vulnerability has been exploited in "morph attacks" in security scenarios. Here, we asked whether the "other-race effect" (ORE)—the human advantage for identifying own- vs. other-race faces—exacerbates morph attack susceptibility for humans. We also asked whether face-identification performance in a deep convolutional neural network (DCNN) is affected by the race of morphed faces. Caucasian (CA) and East-Asian (EA) participants performed a face-identity matching task on pairs of CA and EA face images in two conditions. In the morph condition, different-identity pairs consisted of an image of identity "A" and a 50/50 morph between images of identity "A" and "B". In the baseline condition, morphs of different identities never appeared. As expected, morphs were identified mistakenly more often than original face images. Of primary interest, morph identification was substantially worse for cross-race faces than for own-race faces. Similar to humans, the DCNN performed more accurately for original face images than for morphed image pairs. Notably, the deep network proved substantially more accurate than humans in both cases. The results point to the possibility that DCNNs might be useful for improving face identification accuracy when morphed faces are presented. They also indicate the significance of the race of a face in morph attack susceptibility in applied settings.

mallicksnip@gmail.com .

Additional Key Words and Phrases:

Face morphing; face identification; face matching; deep convolutional neural network; other-race effect

1 INTRODUCTION

Biometrics-based identification and verification systems are deployed widely for a range of security applications, such as border control. **AutomatedBorderControl(ABC)** e-gates commonly employ face-recognition systems to capture a live image of a traveler to automate passport image authentication using face-image matching [8]. If this fails at the live face-recognition stage, the documentation can undergo secondary identity verification by a human border control guard. Accurate identity verification and face matching of travel documentation are critical to determining border-crossing eligibility. This type of face identity matching task with unfamiliar faces is difficult for both human recognizers, such as border control guards, and commercially deployed face-recognition systems [13, 30–32]. The challenges of identity-matching when faces are unfamiliar create a security vulnerability that can be exploited to bypass ABC e-gates through a face-morphing attack.

Face morphs have emerged as a new form of identity fraud [7, 28]. In a face-morphing attack, a morphed image can be created by blending face images of two or more identities. For instance, a morphed image containing a 50/50 average of two identities can be submitted for inclusion in official travel documentation. The live face recognition system may then erroneously verify two different individuals for the same passport image. In an applied setting, criminal actors could morph their faces with a similar-looking noncriminal accomplice and subvert ABC e-gates, due to the resemblance of the live face image to the morph. Face-morphing attacks have been determined to be a feasible method of deceiving face-recognition systems at ABC e-gates [7].

Human behavioral studies also indicate that people are susceptible to morph attacks [20, 31, 32]. For example, in one study [31], when participants were given a face-matching task without being warned about computer-generated morphs, 50/50 morphs were accepted as genuine identities at high rates (68 percent). When participants were warned about the presence of morphed images, the false acceptance rate of 50/50 morphs was reduced significantly (21 percent) [31]. Moreover, training to identify image artifacts that resulted from morph generation (e.g., overlapping hairlines) improved identity-matching performance [32]. The combined effects of morph detection guidance and training led to higher morph identification rates than just morph detection guidance alone [32], although this performance might have been due, in part, to artifact detection [20]. Additionally, individuals who already performed well with distinguishing between two similar-looking faces had better morph identification performance [32].

To address the limitations of morph detection, machine learning–based approaches, including some based on **Deep-ConvolutionalNeural Networks (DCNNs)**, have been leveraged to automate face-morph detection. One early study utilized micro-texture feature extraction and a linear Support Vector Machine (SVM) to determine if a given image was a

morph [26]. In comparison to other feature extraction-based methods such as **Local Binary Patterns-SVM (LBP-SVM)** and **Local Phase Quantisation-SVM (LPQ-SVM)**, the SVM used in this study outperformed previous algorithms. Another study combined the features of two DCNNs, VGG-19 and AlexNet, to explore how transfer learning can impact morph detection in digital and print-scanned images [27]. In comparison to the methods used in previous work [26], such as LBP-SVM and LPQ-SVM, the combined DCNNs performed better on this task. Additionally, a **multiple scales attention convolutional neural network (MSA-CNN)** trained on morph artifacts outperformed other networks like VGG-19 [21] and ResNet18 [38]. Although these algorithms performed relatively well, it is hard to compare performance due to variability in network designs and morph quality.

As morphing software rapidly improves to produce higher quality images, morph recognition could become even more challenging, due to the reduction of obvious artifacts (e.g., overlapping hairlines). In one recent study [20], humans and a VGG-based DCNN were tasked with matching identities in pairs of images that included high-quality morphs. These high-quality morphs were designed to limit artifacts in the morphing process. Morphs were defined as 50/50 combinations of two identities—one of the identities in the morphed image matched the identity of the other face in the pair. Both humans and machines performed poorly at this task.

Individual variations in susceptibility to morph attacks may be impacted further by the difficulties associated with cross-race face identification (e.g., [5], [16], and [36]). The **other-race effect (ORE)** describes the findings that humans recognize faces of their “own” race more accurately than faces of other races [16, 18]. Demographic factors such as the race of a face also affect the performance of face-recognition algorithms such as DCNNs [4, 6, 9, 12]. Although there are no consistent findings on how race impacts algorithm accuracy, there is clear evidence that algorithm performance can be affected by race-based demographic differences (e.g., [4] and [9]). In 2019, algorithms submitted to the **FaceRecognitionVendorTest(FRVT)** showed evidence of demographic differences in face-recognition performance [9]. For example, an algorithm trained on a dataset of immigration application photos had higher false positive rates (erroneous matching of two similar-looking people) for West and East African and East-Asian populations than for Eastern European populations.

Concerns about algorithm performance across variable demographics are exacerbated in the case of morph attacks, especially in airport or border control settings. The high-quality morphs used in [20] included some diverse faces (6 African-American/Black, 16 East Asian, 16 South Asian, and 16 Caucasian). The VGG algorithm used in that study was less accurate at identifying morphed images of Black faces than morphed images of East Asian, South Asian, and White faces. Human identification accuracy as a function of the racial category of the face was not reported. Although the differences in algorithm performance reported by [20] are of interest, the number/balance of faces across race categories was not controlled enough to provide a direct test of the role of race in the identification of morphed images.

The goal of this study was to understand how the ORE influences morph attack susceptibility for both humans and a DCNN algorithm. To directly examine this effect, **East-**

Asian(EA) and **Caucasian(CA)** participants were recruited to complete a face-matching test. The stimuli we used consisted of original images of EA and CA faces and 50-percent morphs of same-race faces (CA-CA and EA-EA morphs). Participants were asked to determine if the faces pictured in image pairs showed the same person or different people. We compared face-matching performance for same-race and other-race morphs and non-morphs (baseline). Participants were unaware that morphed images were present in the test. On the computational side, to compare human and machine performance, a DCNN [29] performed the same task as humans on the same stimuli. To minimize the possibility that morphed images could be perceived as "fake", we presented only the cropped internal regions of the face. The elimination of the external face detail (hair, etc.) also removes identity cues that have been accessible to humans in previous studies. Because most machine-based face-identification algorithms work only on the internal face, this study puts the machine-human comparison on a more equal footing than previous comparisons.

2 HUMAN FACE-IDENTIFICATION EXPERIMENT

2.1 Methods

2.1.1 Design.—The experimental design included three independent variables: participant race (Caucasian, East Asian), face-image race (Caucasian, East Asian), and face-image type (morph, baseline). The latter two varied within-subjects. Accuracy at matching face identity was measured as the area under the receiver operating curve (AUC).

2.1.2 Participants.—A total of 74 students from the University of Texas at Dallas (UTD) participated in this study. The study was conducted virtually, using Microsoft Teams, due to the social-distancing measures put into practice during the COVID-19 pandemic. Students were recruited using the UTD online sign-up system (SONA) and received one course credit as compensation for their participation. All participants were required to be 18 years of age or older, self-identify as Caucasian or East Asian, and have normal or corrected-to-normal vision.

Race and ethnicity eligibility was determined via a recruitment survey generated on Qualtrics [25]. Specifically, the recruitment survey was linked in the experimental description on SONA. The recruitment survey was completed anonymously as follows: The first section of the survey included the consent form for the study. Participants who agreed to participate in the study proceeded with the self-identification question. In the self-identification question, participants were asked which of the following best described their race or ethnic group: (a) East Asian [Thai, Macanese, Japanese, Vietnamese, Chinese, Korean, Taiwanese, Mongolian, and Hong Kong heritage], (b) White/Caucasian [Anglo/European descent], (c) Other Asian, (d) Native American or Alaska Native, (e) Native Hawaiian or Other Pacific Islander, or (f) Other. Participants who selected East Asian or Caucasian proceeded with the final section of the survey. The last section of the survey instructed the participants that the experiment required the installation of MS Teams on a computer (phones and tablets were not permitted for the experiment). Upon agreeing to complete the experiment using MS Teams on a computer, participants were provided with an invitation code that allowed them to enroll in the experiment via SONA. After enrolling,

participants were provided with a link to the SONA experiment corresponding to their demographic group as self-reported in the self-identification question (i.e., a link for either East-Asian participants or Caucasian participants).

Fourteen participants were excluded due to internet connection instability (data collection impediment). The final data included 60 participants. Note that participant recruitment ended when the final data included 30 East-Asian participants (20 female, 10 male, 18–27 years old, average age 20.87) and 30 Caucasian participants (22 female, 8 male, 18–38 years old, average age 22.28). For the survey question “Have you lived in the United States your whole life?”, 21 of 30 EA participants responded “yes” (9 of 30 EA participants responded “no”), and 25 of 30 CA participants responded “yes” (5 of 30 CA participants responded “no”).

A power analysis using PANGEA [37] indicated that a total of 60 participants would be sufficient to obtain a power of 0.839 for a medium effect size ($d = .5$). This power analysis was computed to detect a two-way interaction between face-image race (within-subject, East Asian vs Caucasian) and face-image type (within-subject, Baseline vs Morph).[1]

2.1.3 Stimuli.—A total of 64 face-image pairs were used in this experiment. Each face-image pair was assigned to the morph condition (16 East Asian pairs, 16 Caucasian pairs) or the baseline condition (16 East Asian pairs, 16 Caucasian pairs). The Caucasian and East-Asian groups contained 8 male pairs and 8 female pairs. Both conditions (morph and baseline) included 16 same-identity pairs (two images of the same identity) and 16 different-identity pairs (two images of different identities of the same race, gender, and age group). For each condition, different-identity items were created by randomly pairing same-race and same-gender identities. Additionally, all different-identity image pairs were verified manually to ensure that they contained identities matching in age group. It is important to note that different-identity pairs were not created based on similarity measures.

In the morph condition, different-identity pairs included one unedited image (identity A, image 1) and one 50/50 morph between one image of the same identity (identity A, image 2) and one image of a different identity (identity B, image 1). Same-identity pairs were created using one unedited image (identity A, image 1) and one 50/50 morph between two different images of the same identity (identity A, image 2 and image 3). We used morphs in the same-identity pairs to support the Signal Detection Model measures, which require both same- and different-identity pairs in each condition. This ensured also that the performance observed in the morph condition was derived from people’s ability to distinguish same- and different-identity pairs, rather than morphed and non-morphed images. In the baseline condition, same-identity pairs were created using one unedited image (identity A, image 1) and one cropped image of the same identity (identity A, image 2). Different-identity pairs included one unedited image (identity A, image 1) and one cropped image of a different

[1]Note that the design of the power analysis conducted prior to data collection was inaccurate to estimate the sample size required to detect a three-way interaction. A secondary analysis was computed to detect a three-way interaction between participant race (between-subject, East Asian vs Caucasian), face-image race (within-subject, East Asian vs Caucasian) and face-image type (within-subject, Baseline vs Morph). Results confirm that a total of 60 participants (30 per participant race group) is sufficient to obtain a power of 0.839 for a medium effect size ($d = 0.5$).

person (identity B, image 1). See Figure A1 for an example of the stimulus pairs for each condition.

All morphed images were cropped around the face to minimize morph artifacts. The algorithm tested in this experiment operates on the internal face. To do a true machine comparison, these morph artifacts were excluded so the observers were limited to landmarks that a facial recognition system uses (eyes, nose, facial structure, etc.). Image-morphing software cannot adequately account for hair across different images. When including hair in a morphed image, the hair either will become blurred or must be added/rendered after the fact so that it appears photo-realistic.

Images were selected from the Notre Dame Database [33] and showed faces viewed from the front with neutral expressions. The race and gender of the faces in each pair were balanced across the conditions. In each face-image pair, the unedited images consisted of images captured in an uncontrolled illumination setting. All image manipulations (morphing and cropping) were executed on images captured under controlled illumination and performed using the Face Morpher Github repository [39]. Additionally, all morphed images underwent further editing with Photoshop and Gimp to remove artifacts (e.g., second irises, smooth appearance, overlapping noses, etc.). Following morphing and cropping, images underwent sharpening in Photoshop to reduce blurred complexions.

2.1.4 Remote Testing Protocol.—

In order to comply with the COVID-19 social-distancing requirements, human data collection was carried out virtually. The experiment was conducted online using the remote-control features available on Microsoft Teams. Participants were required to complete the experiment on a personal computer. Other devices such as phones or tablets were not permitted for study participation. Aspects pertaining to the participants' environment (e.g., lighting, noise, distraction, etc.) were not controlled. All human data were stored locally on the experimenter's computer. The experiment was conducted using PsychoPy v1.84.2 [22]. All participants used Qualtrics survey software to complete the Self-identification survey.

2.2 Procedure

2.2.1 Face-Matching Task.—

All eligible participants received an invitation link to participate in a conference call with the experimenter. The face-matching task was administered virtually using Microsoft Teams. The experiment was conducted locally on the experimenter's computer. During the experimental session, the subject was given permission to view the experimenters' screen (via screen sharing) and control the experimenters' mouse and keyboard remotely. After giving informed control, the participant proceeded with the face-matching task.

The face-matching test included a total of 64 trials (Figure A1(B)). The face-image pairs in each condition (face-image race, face-image type) were presented in a randomized order. Information pertaining to the experimental conditions (face-image race and face-image type) was not revealed explicitly. On each trial, a face-image pair was presented on the screen. The participants were instructed to determine whether the two images were of the same identity or different identities. Responses were collected using a 5-point certainty scale (1: Sure they

are the same; 2: Think they are the same; 3: Do not know; 4: Think they are not the same; 5: Sure they are not the same). Participants did not have a response time limit and the stimuli remained on the screen until a response was entered.

After completing the face-matching task, the subject was instructed to complete a short demographic survey (see Appendix, Figure A.1).[2]

2.3 Results

Data were analyzed using a 2 (participant race: East Asian vs Caucasian) × 2 (face-image race: East Asian vs Caucasian) × 2 (face-image type: Baseline vs Morph) mixed-model ANOVA. Face-image race and face-image type were submitted as within-subjects factors and participant race was submitted as a between-subjects factor. The dependent variable (face-matching accuracy) was measured as the AUC. The AUC was computed based on a construction of the **receiver operating characteristic (ROC)** curve, using the standard method described in [15] for Likert scale data. This uses rating scale points from the Likert as criteria at which hit and false alarm rates can be computed and integrated for the creation of the ROC, thereby supporting the computation of an AUC.

In what follows, we report multiple interactions, including a three-factor interaction among participant race, face-image race, and face-image type. For clarity and completeness, we begin with lower-order effects. As always, interpretations of lower-order effects are tentative and subject to change in the presence of higher-order interactions. As expected, participants performed more accurately for the baseline image pairs ($M = 0.841$, $SE = 0.012$, 95% CI [0.818, 0.865]) than the morphed image pairs ($M = 0.725$, $SE = 0.011$, 95% CI [0.703, 0.746]) (see Figure A2). Specifically, there was a main effect of face-image type ($F(1,58) = 77.283$, $MSe = 0.011$, $p < .001$, $\eta_p^2 = 0.571$). No other main effects were significant.

There was a significant two-way interaction between participant race and face-image race. When averaged across the two image types (morphed and baseline), Caucasian participants were more accurate at identifying Caucasian face pairs ($M = 0.811$, $SE = 0.015$, 95% CI [0.780, 0.842]) than East-Asian face pairs ($M = 0.773$, $SE = 0.016$, 95% CI [0.741, 0.806]), and East-Asian participants performed similarly for East-Asian face pairs ($M = 0.781$, $SE = 0.016$, 95% CI [0.748, 0.813]) and Caucasian face pairs ($M = 0.767$, $SE = 0.015$, 95% CI [0.736,0.798]). Although this would seem to suggest that only Caucasian participants show the ORE, an interpretation of the two-way analysis must await an analysis of the three-way interaction. No other two-factor interactions were significant.

Of primary interest for this study, there was a three-way interaction between participant race, face-image race, and face-image type ($F(1,58) = 4.49$, $MSe = 0.0073$, $p = 0.038$, $\eta_p^2 = 0.07$). Figure A2 shows that both the East-Asian and Caucasian participants fared equally well on East-Asian and Caucasian face pairs in the baseline condition. In the morph condition, however, there was an ORE such that East Asians performed more accurately on the East-Asian morph pairs and Caucasians performed more accurately on the Caucasian

[2]Participant selection (eligibility) was not determined by the recruitment survey

morph pairs. In other words, the two-factor interaction we found between the race of the participants and face was driven by the difficulties participants had with other-race morphs.

In summary, the human experiment replicates the well-documented difficulties people have in matching face identities with morphed stimuli [20, 31]. Notably, we did not find an ORE in the baseline condition–only in the morph condition, as was evident from the pattern of means in the three-factor interaction. It is unclear why we did not find a standard ORE in the baseline condition. One possible factor in the lack of an ORE in the baseline condition is that the local population of students in Dallas is highly diverse and so students would be in constant contact with people of many races.[3] A second factor is that, because the majority of both East Asian (70%) and Caucasian (83.33 %) participants sampled in the present study reported living in the United States their whole life, both groups of participants may have experienced similar "face diets" and may therefore perform similarly when identifying face images of different races [19]. A third factor is that performance in the baseline condition was quite accurate. It may be that the ORE is most easily seen in more challenging conditions, such as with morphs. The finding of a crossover interaction in the morph, but not the baseline condition, is consistent with this interpretation. However, it is not possible to know for sure why we did not find the classic ORE in the baseline condition. Notwithstanding, the results indicate the additional challenge of face identification with other-race morphs.

Additionally, accuracy for same-identity pairs was measured for all conditions (face-image type and face-image race). Accuracy was determined by the proportion of correct responses ("Sure they are the same" or "Think they are the same") endorsed to same-identity items. The data were submitted to a 2 (participant race: East Asian vs Caucasian) × 2 (face-image race: East Asian vs Caucasian) × 2 (face-image type: Baseline vs Morph) mixed-model ANOVA. Face-image race and face-image type were submitted as within-subjects factors and participant race was submitted as a between-subjects factor. Accuracy (proportion correct) was treated as the dependent variable. The results did not reveal any significant effect or interaction. Specifically, accuracy was not significantly different for same-identity image pairs in either face-image type condition (Baseline vs Morph) or face-image race condition (East Asian vs Caucasian).

3 DCNN EXPERIMENTS

3.1 Methods

3.1.1 Network Architecture.—We used a recent high-performing (cf. [17]) DCNN [29] based on the ResNet-101 architecture [11]. The network contains 101 layers and was trained using the Universe face dataset [1, 29]. When introduced, the dataset was not named [1], but it has been referred to in subsequent publications as the "Universe" dataset (e.g., [29]). It is a compilation of three smaller face-image datasets (UMDFaces [2], UMDVideos [1], and MS-Celeb-1M [10]), with no additional images added beyond those in the original three datasets. However, a semi-automated hierarchical clustering method [14] was used to remove incorrectly labeled images from the MS-Celeb-1M dataset. In total, the Universe

[3] https://ospa.utdallas.edu/common-data-set/

dataset contains 5,714,444 images of 58,020 identities. The images within this dataset are sampled to include considerable variation in image parameters (e.g., pose, illumination, resolution, etc.) across face images of a given identity [1, 3]. However, the demographic information of the identities comprising the dataset is not known. During training, the network used Crystal Loss with the alpha parameter set to 50. Skip connections are used throughout the 101-layered network to retain the amplitude of the error signal. After training was complete, the final fully-connected layer of the network was removed and the output from the penultimate layer (containing 512 units) was used to generate identity descriptors.

We chose this network because it has been used in previous human-machine comparisons with both expert professional forensic face examiners and untrained participants [24]. The network performed more accurately than untrained participants and performed at the level of professional forensic face examiners on a challenging face-identification task with a majority of CA faces. In addition, this network has been shown to maintain high accuracy even across considerable variability in pose, illumination, and expression [17]. The network has been used also to test performance differences between CA and EA faces [4]. In the multi-race tests, overall network performance (AUC) was roughly comparable for EA and CA faces. However, at the low false alarm rates commonly used in security applications, CA faces were identified more accurately than EA faces. Finally, the results produced by DCNNs based on a ResNet-101 architecture have been shown to possess high similarity to perceptual responses recorded in the human brain, as measured using the “BrainScore” metric [34, 35]. Combined, all of these factors contributed to our selection of the network used in the present study as an appropriate network for our research.

3.1.2 Procedure.—

Each of the face images used in the human experiment was processed through the DCNN to produce face-image descriptors. All face images were successfully detected and processed by the network regardless of whether the image was manipulated (i.e., edited or morphed).

For each image pair in the human experiment, the cosine similarity (i.e., normalized dot product) between image descriptors was computed. Higher similarity scores were assumed to indicate a higher likelihood that the images showed the same identity. To assess network accuracy, an AUC was computed from distributions of similarity scores for the same- and different-identity pairs in each condition.

3.2 Results

For image pairs in the baseline (i.e., non-morphed) condition, DCNN face-identification accuracy was perfect (AUC = 1.0, see Figure A3 left). For image pairs in the morph condition, DCNN face-identification accuracy was substantially lower (AUC = 0.891, see Figure A3 right). The decrease in DCNN identification accuracy was more pronounced for Caucasian images (baseline AUC = 1.0; morph AUC = 0.859) than East-Asian images (baseline AUC = 1.0; morph AUC = 0.922).

In summary, the DCNN performed perfectly on the baseline image pairs and less accurately on the morphed image pairs. Furthermore, the DCNN showed an accuracy advantage for EA over CA morph pairs.

4 SUMMARY: HUMAN AND MACHINE PERFORMANCE

Both humans and machines showed an advantage for recognizing baseline over morphed images, consistent with previous studies [31]. Human participants showed an ORE in the morph condition, but not in the baseline condition. Although the DCNN was not tested for a cross-over ORE, the performance of the network was analyzed as a function of the race of the morphed faces. Overall, the performance of the DCNN surpassed humans on both baseline and morphed image pairs.

5 DISCUSSION

The principles underlying the ORE in humans have been well studied [16, 18]. In response to the emerging security threat posed by face morphs, we analyzed the influence of participant and stimulus race on morph attack susceptibility in humans and a DCNN. The present findings expand our understanding of how participant and stimulus race combined influence face identification in a morph-attack scenario. Our human behavioral results demonstrate that these factors combine to exacerbate the problem of face identification when images are morphed. Specifically, morphs pose a particularly strong challenge to an observer of a different race than the morph. The DCNN used in this study also performed more accurately than human participants in all cases. Thus, despite its reduced performance for morphed image pairs, and the differences in accuracy for EA and CA faces, the DCNN was always the more accurate face identification “system”. The findings from this study have significant implications for understanding how race could bias human and algorithmic decision-making in border control scenarios. We consider each of these implications in turn.

This is the first study to assess systematically the role of participant race and face-image race on morph identification. This was accomplished by conducting a complete cross-over design. Thus, Caucasian and East-Asian participants were tested on CA and EA face-image pairs. The present study provides evidence that for morphed images, morph-attack susceptibility is increased when the observer and face are of different races. In addition to the use of a cross-over design, the present study provides a more direct test of the role of race in the identification of morphs for humans and machines. First, we controlled for the possibility that people could detect artifacts in morphed images by cropping the faces to include only the internal face. This also made for a more equitable comparison between the DCNN, which works only on the internal face, and humans. Second, we removed face-image artifacts (e.g., overlapping irises, smooth complexions) that were introduced during the morphing process. Additionally, this study used morphed image pairs for same-identity comparisons to ensure a common ground truth between same- and different-identity comparisons.

The finding that DCNN performance was more accurate for East-Asian than Caucasian morphed-image pairs underscores the unpredictability of algorithms for faces of different races. Although it is clear that the performance of DCNNs are affected by demographics, the source of these effects is less clear and remains an active and open area of research [4, 9]. Previous studies have indicated that algorithms that originated in China tended to have lower false positive rates on East-Asian faces, although it is not clear whether this

difference resulted from training, optimization, or some other unknown parameter of the algorithms [9]. This type of race bias has been demonstrated also in pre-DCNN algorithms. Earlier algorithms developed in Western countries (e.g., France, Germany, the United States) performed more accurately on Caucasian faces, whereas algorithms developed in East Asian countries (e.g., China, Japan, Korea) performed more accurately on East-Asian faces. Again, however, the source of the effects is not known [23]. In the current study, differences in the performance of the DCNN on EA and CA faces could likewise have resulted from a variety of factors, including imbalances in the training set composition (age, race, etc.), as well as image quality differences [4, 9]. Notwithstanding the demographic effects, the DCNN used in this study fared far better than humans on both baseline and morphed images.

This study lays the groundwork to conduct future assessments for how the observer race and face race could affect morph identification across multiple races. One limitation of this experiment is the consideration of only two racial groups, a limit that can be overcome in future work by expanding the range of racial diversity of participants and face images. Concomitantly, there is a wide diversity of demographic effects in DCNNs [9]. Thus, it is incumbent on algorithm users to carefully test and validate the performance of specific algorithms for morphs of different race faces intended for particular applications (e.g., airports in different locations around the world).

The results show that the ORE exacerbates the difficulties associated with morph identification. As fraudsters find new and creative ways of bypassing ABC e-gates, this study elucidates a path forward to mitigate the incidence of morph attacks by investigating how race influences humans and algorithms. The findings have implications for national and international security measures and underscore the complexities of morphed face identification by humans and algorithms.

Acknowledgments

Funding provided by National Eye Institute Grant R01EY029692-04 to AOT and CDC..

A: APPENDIX

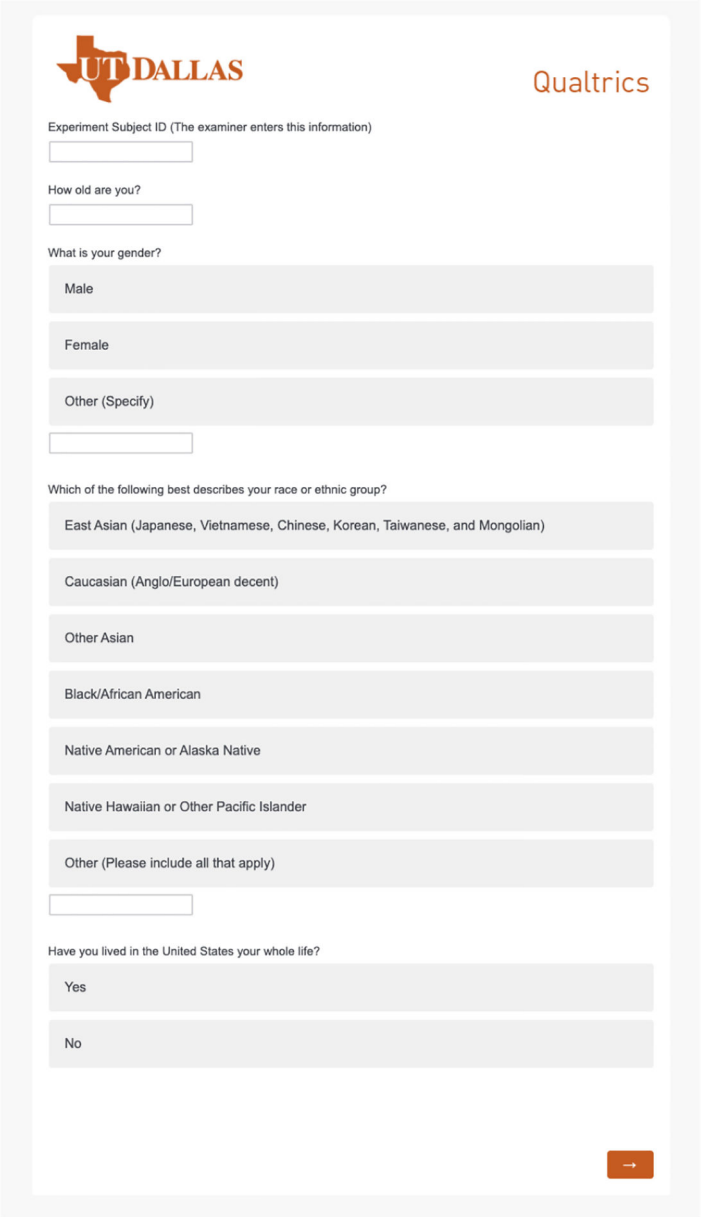

Fig. A.1.
Demographic survey.

REFERENCES

[1]. Bansal Ankan, Castillo Carlos, Ranjan Rajeev, and Chellappa Rama. 2017. The do's and don'ts for CNN-based face verification. In Proceedings of the IEEE International Conference on Computer Vision Workshops. 2545–2554.

[2]. Bansal Ankan, Nanduri Anirudh, Castillo Carlos D., Ranjan Rajeev, and Chellappa Rama. 2017. UMDfaces: An annotated face dataset for training deep networks. In Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB'17). IEEE, 464–473.

[3]. Bansal Ankan, Ranjan Rajeev, Castillo Carlos D., and Chellappa Rama. 2018. Deep features for recognizing disguised faces in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 10–16.

[4]. Cavazos Jacqueline G., Phillips P. Jonathon, Castillo Carlos D., and O'Toole Alice J.. 2020. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science 3, 1 (2020), 101–111. [PubMed: 33585821]

[5]. Chiroro Patrick and Valentine Tim. 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 (1995), 879–894.

[6]. Drozdowski Pawel, Rathgeb Christian, Dantcheva Antitza, Damer Naser, and Busch Christoph. 2020. Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society 1, 2 (2020), 89–103.

[7]. Ferrara Matteo, Franco Annalisa, and Maltoni Davide. 2014. The magic passport. In Proceedings of the IEEE International Joint Conference on Biometrics. IEEE, 1–7.

[8]. Frontex. 2015. Best Practice Technical Guidelines for Automated Border Control (ABC) Systems.

[9]. Grother Patrick J., Ngan Mei L., and Hanaoka Kayee K.. 2019. Face recognition vendor test part 3: Demographic effects. (2019).

[10]. Guo Yandong, Zhang Lei, Hu Yuxiao, He Xiaodong, and Gao Jianfeng. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Proceedings of the 2016 14th European Conference on Computer Vision, Part III (ECCV 2016) (Amsterdam, The Netherlands, Oct. 11 – 14, 2016). Springer, 87–102.

[11]. He Kaiming, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 770–778.

[12]. Klare Brendan F., Burge Mark J., Klontz Joshua C., Vorder Bruegge Richard W., and Jain Anil K.. 2012. Face recognition performance: Role of demographic information. IEEE Transactions on Information Forensics and Security 7, 6 (2012), 1789–1801.

[13]. Kramer Robin S. S., Mohamed Sophie, and Hardy Sarah C.. 2019. Unfamiliar face matching with driving licence and passport photographs. Perception 48, 2 (2019), 175–184. [PubMed: 30799729]

[14]. Lin Wei-An, Chen Jun-Cheng, and Chellappa Rama. 2017. A Proximity-Aware Hierarchical Clustering of Faces. arXiv:1703.04835 [cs.CV]

[15]. Macmillan Neil A. and Creelman C. Douglas. 2004. Detection Theory: A User's Guide. Psychology Press.

[16]. Malpass Roy S. and Kravitz Jerome. 1969. Recognition for faces of own and other race. Journal of Personality and Social Psychology 13, 4 (1969), 330. [PubMed: 5359231]

[17]. Maze Brianna, Adams Jocelyn C., Duncan James A., Kalka Nathan D., Miller Tim, Otto Charles, Jain Anil K., Niggel W. Tyler, Anderson Janet, Cheney Jordan, and Grother Patrick. 2018. IARPA janus benchmark - C: Face dataset and protocol. In Proceedings of the 2018 International Conference on Biometrics (ICB) (2018), 158–165.

[18]. Meissner Christian A. and Brigham John C.. 2001. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. Psychology, Public Policy, and Law 7, 1 (2001), 3.

[19]. Mousavi Seyed Mortezaand Oruc Ipek. 2020. Tuning of face expertise with a racially heterogeneous face-diet. Visual Cognition 28, 9 (2020), 523–539.

[20]. Nightingale Sophie J., Agarwal Shruti, and Farid Hany. 2021. Perceptual and computational detection of face morphing. Journal of Vision 21, 3 (2021), 4–4.

[21]. Parkhi Omkar M., Vedaldi Andrea, Zisserman Andrew. 2015. Deep face recognition. In BMVC, Vol.1.6.

[22]. Peirce Jonathan W.. 2007. PsychoPy–psychophysics software in Python. Journal of Neuroscience Methods 162, 1–2 (2007), 8–13. [PubMed: 17254636]

[23]. Jonathon Phillips P, Jiang Fang, Narvekar Abhijit, Ayyad Julianne, and O'Toole Alice J.. 2011. An other-race effect for face recognition algorithms. ACM Transactions on Applied Perception (TAP) 8, 2 (2011), 1–11.

[24]. Jonathon Phillips P, Yates Amy N., Hu Ying, Hahn Carina A., Noyes Eilidh, Jackson Kelsey, Cavazos Jacqueline G., Jeckeln Géraldine, Ranjan Rajeev, Sankaranarayanan Swami, Chen Jun-Cheng, Castillo Carlos D., Chellappa Rama, White David, and O'Toole Alice J.. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences (2018), 201721355.

[25]. Qualtrics. 2019. Qualtrics,2019. https://www.qualtrics.com/

[26]. Raghavendra R, Raja Kiran, and Busch Christoph. 2016. Detecting morphed face images. 10.1109/BTAS.2016.7791169

[27]. Raghavendra R, Raja Kiran B., Venkatesh Sushma, and Busch Christoph. 2017. Transferable deep-CNN features for detecting digital and print-scanned morphed face images. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'17). 1822–1830. 10.1109/CVPRW.2017.228

[28]. Raghavendra RB, Raja Kiran, Venkatesh Sushma, and Busch Christoph. 2017. Transferable deep-CNN features for detecting digital and print-scanned morphed face images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[29]. Ranjan Rajeev, Bansal Ankan, Xu Hongyu, Sankaranarayanan Swami, Chen Jun-Cheng, Castillo Carlos D., and Chellappa Rama. 2018. Crystal loss and quality pooling for unconstrained face verification and recognition. arXiv preprint arXiv:1804.01159 (2018).

[30]. Ritchie Kay L., Mireku Michael O., and Kramer Robin S. S.. 2020. Face averages and multiple images in a live matching task. British Journal of Psychology 111, 1 (2020), 92–102. [PubMed: 30945267]

[31]. Robertson David J., Kramer Robin S. S., and Burton A. Mike. 2017. Fraudulent ID using face morphs: Experiments on human and automatic recognition. PLoS One 12, 3 (2017), e0173319. [PubMed: 28328928]

[32]. Robertson David J., Mungall Andrew, Watson Derrick G., Wade Kimberley A., Nightingale Sophie J., and Butler Stephen. 2018. Detecting morphed passport photos: A training and individual differences approach. Cognitive Research: Principles and Implications 3, 1 (2018), 1–11. [PubMed: 29399620]

[33]. Schott Cathy L. and Sharpe Matthew. 2010. FRVT 2006 and ICE 2006 large-scale experimental results. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 5 (2010), 831. [PubMed: 20299708]

[34]. Schrimpf Martin, Kubilius Jonas, Hong Ha, Majaj Najib J., Rajalingham Rishi, Issa Elias B., Kar Kohitij, Bashivan Pouya, Jonathan Prescott-Roy Kailyn Schmidt, Yamins Daniel L. K., and DiCarlo James J.. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? BioRxiv (2018), 407007.

[35]. Schrimpf Martin, Kubilius Jonas, Lee Michael J., Ratan Murty N. Apurva, Ajemian Robert, and DiCarlo James J.. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. Neuron (2020). https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X

[36]. Walker Pamela M. and Tanaka James W.. 2003. An encoding advantage for own-race versus other-race faces. Perception 32, 9 (2003), 1117–1125. [PubMed: 14651324]

[37]. Westfall Jacob. [n.d.]. PANGEA PANGEA : Power ANalysis for GEneral Anova designs. https://api.semanticscholar.org/CorpusID: 43131842

[38]. Zhang Le-Bing, Cai Juan, Peng Fei, and Min Long. 2022. MSA-CNN: Face Morphing Detection via a Multiple Scales Attention Convolutional Neural Network. 17–31. 10.1007/978-3-030-95398-0_2

[39]. Quek Alyssa and Morpher Face. [online] Available: https://github.com/alyssaq/face.morpher

CCS Concepts:

• **Applied computing → Psychology; • Security and privacy → Biometrics**;

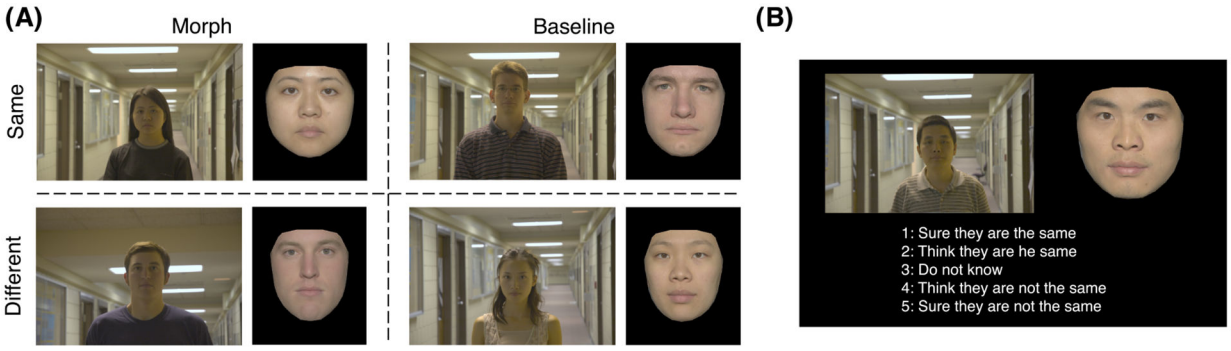

Fig. A1.

(A) Morph condition: Face-image pairs included one unedited image and one cropped 50/50 face morph. The face morphs were created by blending two images of the same identity (n = 16) or blending two images of different identities (n = 16). Baseline condition: Face-image pairs included one unedited image and one cropped image of the same identity (n = 16) or one cropped image of a different identity (n = 16). (B) Example of a face-matching trial.

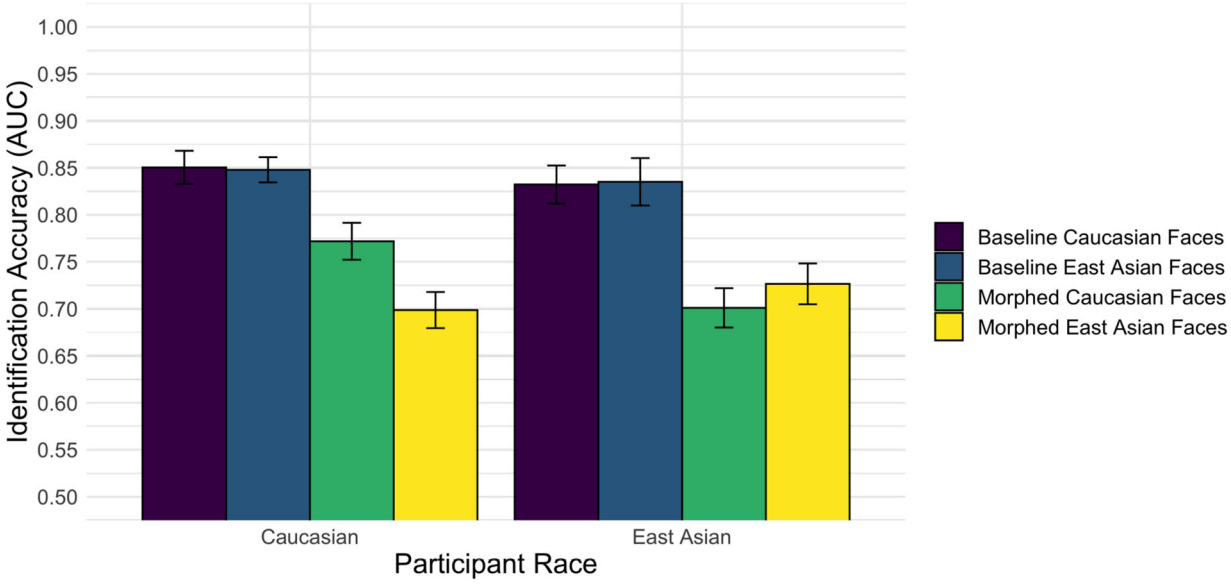

Fig. A2.

Face-identity matching results. Performance was more accurate for baseline pairs (purple and blue bars) than morph pairs (green and yellow bars). Notably, other-race morph pairs proved especially difficult for both East-Asian and Caucasian participants (green and yellow bars).

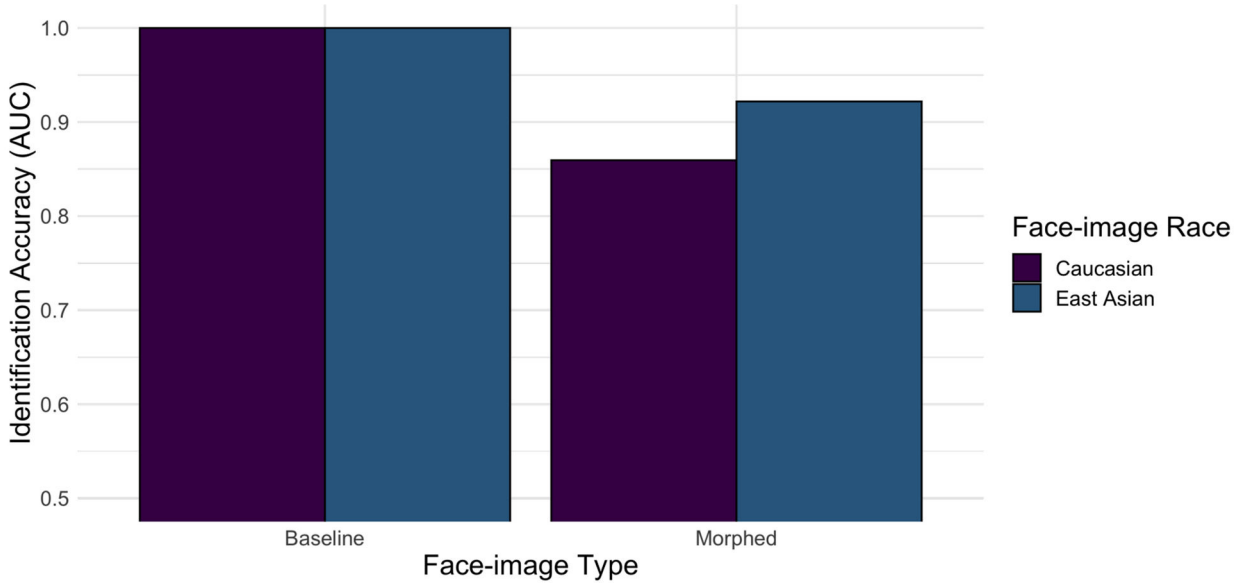

Fig. A3.

DCNN-based identification accuracy for Caucasian and East-Asian face-image pairs. Accuracy was lower for morphed image pairs in comparison to baseline image pairs, and lower for Caucasian morphed image pairs than East-Asian morphed image pairs.