Human ability to match synthesised faces to their constituent faces

[redacted]

This thesis is submitted in partial fulfilment
of the Honours Degree of Bachelor of Psychology

School of Psychology
The University of Adelaide
October 2017

Word Count: 9239

Table of Contents

List of Figures

List of Tables

Abstract

Currently, facial recognition systems are used widely within various government agencies as a means of identity verification, and these systems involve a human operator in the final decision making process. Previous studies have shown that both face recognition systems and humans are vulnerable to variables that may impede the face matching process, including the use of morphed images, which are created by digitally combining multiple constituent faces into a new face. Therefore, the current study aims to further investigate how the usage of different types of morphed faces can affect human face matching performance. Participants (*N* = 51) from the University of Adelaide and the general public completed 112 computer-based one-to-many face matching trials in a repeated measures design. The type of morphed face used as the target image varied for each trial, and was either made from 2, 8 or 16 constituent faces, or was a control non-morphed face. Results indicated that the usage of 8-Image morphs resulted in significantly higher accuracy and confidence, as well as faster response latency. Future research could be conducted using morphs made from similar faces, and employ multidimensional scaling methods to map the morphs and their constituent faces in face space.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and to the best of my knowledge, contains no materials previously published except where due reference is made. I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

October 2017

Acknowledgements

Firstly, I would like to thank my internal supervisor Dr Carolyn Semmler at the University of Adelaide for her guidance and assistance throughout the year. I appreciate your reassurance, the short-notice meetings you were able to accommodate, and always keeping me in the right direction. I would also like to extend a very big thank you to my external supervisor Dr Rebecca Heyer at DST Group. I cannot thank you enough for your endless help, the constructive responses to my countless emails, and the time and effort you invested in this project.

I would also like to extend a big thank you to my parents, Maria and Paul, and my brother Zachary for their unconditional support, positivity, and reassurance throughout even the most stressful of times. I would not have been able to get through this year without them. Finally, I would like to thank my friends, both in-and-outside of university, for making this year easier to manage.

CHAPTER 1

Introduction

1.1 Rationale

In many government agencies worldwide, facial recognition systems are used in the process of identification, including for detecting duplicate applications for identity documents by the same person (White, Dunn, Schmid, & Kemp, 2015). However, these systems are not fool-proof. Unlike under laboratory conditions, identification accuracy in the real-world can be poor, especially in situations where the image capture conditions were not ideal (Phillips & O'Toole, 2014). Therefore, to overcome this shortcoming, many applications require a human component in the face matching process, providing a candidate list of the highest matching images for human review. The human operator is then required to make a decision on whether the face they are interested in (i.e. the ‘target image’) corresponds to any of the faces in that candidate list (Graves et al., 2011). In Australia, facial recognition systems are used mainly for identity verification. It is estimated that nearly five percent of Australians will experience identity theft each year, which makes it more common that assaults or other forms of thefts (Attorney-General's Department, n.d.). There is also worry that fraudulent identities can be used to facilitate acts of domestic terrorism and other crimes, making the verification of one’s identity a key issue for national security.

One potential way of creating fraudulent identifications is to use manipulated facial images, in this case the usage of facial morphs. In this scenario, graphical morphing programs can be used to generate a sequence of images, containing the blending of multiple faces along a continuum (Robertson, Kramer, & Burton, 2017). At the central point of the continuum, the morphed face would contain equal amounts of each face. As this is a new area of interest, not much is known regarding how the introduction of a morphed face in a face matching task will impact on performance.

1.2 Unfamiliar Face Matching

Familiar faces can be defined as faces that are famous or personally familiar to a person, or faces in a study after an extensive learning phase (Johnston & Edmonds, 2009). *Unfamiliar faces*, on the other hand, are faces that a person has never seen before, or have been seen for an insufficient period of time for the brain to construct a stable memory representation (Hancock, Bruce, & Burton, 2000). Both types of faces are parsed by different underlying mechanisms (Megreya & Burton, 2006). However, this study is only interested in unfamiliar face matching.

1.2.1 Processing of unfamiliar faces. In the literature, the dual-code view is the predominant theory for unfamiliar face processing, where it states that both configural and featural processing play an important role (Cabeza & Kato, 2000). Featural processing uses the information encoded within the facial features themselves (Schwaninger, Lobmaier, & Collishaw, 2002). On the other hand, configural processing is used to analyse the spatial relationship between facial features, and combining the individual features together into a gestalt (Carbon, 2011; Maurer, Grand, & Mondloch, 2002). However, the contribution of both forms of processing is not equal. There is growing evidence that unfamiliar faces are processed in a more object-like fashion, lending support to the idea that featural processing is more commonly used when looking at unfamiliar faces (Johnston & Edmonds, 2009).

Towler, White, and Kemp (2017) have shown that a feature-to-feature image comparison strategy, which has its roots in research involving featural processing, can improve face matching performance. When participants were asked to rate the target images on their similarities, their accuracy in the subsequent face matching task were improved, compared to rating the faces on image quality and perceived personality traits. Although the effect was limited to match trials only, signal detection methods showed that the results reflected an improvement in identity discrimination. This is the reason why featural processing is the recommended method of facial comparison for most government agencies (FISWG, 2012).

1.2.2 Matching unfamiliar faces. In general, people are often quite poor at performing tasks involving unfamiliar faces. In the landmark study by Bruce et al. (1999), performance ranged from 50% to 96% accuracy, and this has been replicated by multiple studies (Megreya & Burton, 2008; White, Kemp, Jenkins, Matheson, & Burton, 2014) and even persists with different presentation modalities. For example, performance is still less than optimal when participants were asked to match a live person and a photograph (Megreya & Burton, 2008). However, it is vital to note that face matching tasks have large performance variance due to individual differences (Kemp, Towell, & Pike, 1997). This could be due to individual variations in perception and memory capabilities, as well as total cerebral volume and rational-experiential thinking styles (Calic, 2013; Megreya & Burton, 2006; Schretlen, Pearlson, Anthony, & Yates, 2001).

1.2.3 Challenges to face matching performance. People are much more sensitive to external influences when matching unfamiliar faces, including age, differences in viewpoint, expression and lightning, and occlusion. Ageing is one of the most widespread issues faced by users of facial recognition systems, and this includes changes to the facial texture (like the appearance of wrinkles), shape and facial hair (Ling, Soatto, Ramanathan, & Jacobs, 2007). Studies have found that commercial facial recognition algorithms fail to match the subjects in the images as the age gap between images increases, and this effect is expected to be present as well in human operators (Albert, Sethuram, & Ricanek, 2011; Ling et al., 2007).

Additionally, changes in viewpoint, lighting and facial expression can affect face matching performance. Regarding viewpoint changes, as the face is rotated towards the profile, there is a trade-off in available facial information. Less information is available regarding the spatial relationship between features, and the viewer is able to determine the angle of the forehead and nose (Van der Linde & Watson, 2010). Therefore, the featural information one can obtain from a forward-facing face and a profile face is significantly different. This is further influenced by changes in lighting, where different performance was found when comparing top-lit or bottom-lit faces (Hill & Bruce, 1996). Different facial expressions can also pose a threat, as expressions, like disgust or anger, can reduce accuracy and increase reaction time in a face matching task (W. Chen, Lander, & Liu, 2011).

Furthermore, in real-life scenarios, there could also be interference from the facial features themselves. This could include changes in hairstyle, moles, scars and other paraphernalia. Studies have found that the manipulation of facial features increases the false acceptance rate, especially in mismatch trials where both images contain identical spectacles, hairstyle or location of moles (Wirth & Carbon, 2017). In another scenario, certain facial features can also be occluded by accessories, like sunglasses or scarves (Mi, Li, Li, Liu, & Liu, 2016). This obstructs the facial features necessary for identification, which can reduce performance. The decrease in performance is most noticeable in ‘mixed’ scenarios, where the presence of the paraphernalia is difference between the target image and the trial image (Kramer & Ritchie, 2016).

This means that, in theory, matching faces in real-life scenarios is made even more complicated by interference from the above factors. Individuals who have to perform this task as part of their occupational duties, like border protection officials, are also under high cognitive load and experience time pressures, which will further affect performance (McCaffery & Burton, 2016; Wirth & Carbon, 2017). Furthermore, this effect cannot be mitigated by experience. Studies have shown that expert observers, who are those that have received training in facial image comparison as part of their job, are not better than untrained participants in a face matching task (Burton, Wilson, Cowan, & Bruce, 1999; White, Kemp, Jenkins, Matheson, et al., 2014). However, more recent studies have indicated that face matching experts perform better than untrained students, suggesting that the training received has been improved (White, Phillips, Hahn, Hill, & O'Toole, 2015; Wirth & Carbon, 2017). Even so, the expert group still had a surprisingly high false acceptance rate (Wirth & Carbon, 2017). This has alarming implications for national security, as it suggests that it may be possible to enter a country with falsified identity documents.

1.3. Morphed Faces

1.3.1 Background: Face Space. The psychological face space is a theoretical construct that aims to explain the mechanisms of both familiar and unfamiliar face matching. It can be defined as an internal, multidimensional space, where individual faces are stored as single points that vary according to specific dimensions (Nestor, Plaut, & Behrmann, 2013). These dimensions can be parameters like the size of the face, the distance between features, age, or even masculinity (Valentine, Lewis, & Hills, 2016). According to Valentine (1991), there are two different models that explain how faces are encoded within the face-space framework. The norm-based model suggests that faces are encoded based on their deviation from a specific prototype (i.e. a norm). Therefore, distinctive faces are located further away from the norm, while typical faces are located closer to the norm. On the other hand, the exemplar-based model argues that faces are encoded in the face space without referencing any central prototypes. In this scenario, distinctive faces are located in areas of low face representation density. Conversely, typical faces are located near the centre of the distribution, where there is a high face representation density.

An example of how face space literature affects face matching performance is through the use of caricatures, which are images where identifiable facial features have been exaggerated. Studies have shown that systemic caricaturing can increase the distinctiveness of a face (McIntyre, Hancock, Kittler, & Langton, 2013). This makes it easier to discriminate the faces in mismatch trials, as shown by a reduction in false positives. Therefore, morphed faces would likely have the opposite effect, as the distinctiveness of facial features are averaged in the process of morphing two separate images. Hence, morphed images would logically result in a reduction in performance, especially in mismatch trials.

1.3.2 Facial Recognition Algorithm Performance. Currently, there is a limited number of papers that have examined the effect of morphed faces on facial recognition algorithms. A study by Ferrara, Franco, and Maltoni (2014) examined the impact of morphs in an automated facial recognition system. When the system was tested with 10 morphed images (5 males, 5 females), the system accepted the morphed images as similar to both test images, which were alternate images of the people used to create the morph. This attack was even found successful when morphs of images of different genders, or even images from more than 2 different people were used. Therefore, they concluded that automated systems, and in extension human operators as well, are sensitive to these kinds of image alterations.

1.3.3 Human Face Matching Performance. So far, only one study has examined human face matching performance with the inclusion of morphed faces. Robertson et al. (2017) looked at the impact of morphs in a one-to-one face matching task. They found that when a 50% morph was used (made with only two constituent faces) and the participants were blinded towards the true purpose of the study, there was a significantly high false acceptance rate, compared to using the original images. It is important to note that the morphed images used in the Robertson et al. (2017) experiments were compared in 2 different sizes, with the comparison image presented as the face only – with no body visible. The results of the experiments indicate that, similar to automated systems, people are also vulnerable to being deceived by morphs. However, as some human operators are required to compare an image with a candidate list of potential matches (i.e. a one-to-many face matching task), there is still a need to further investigate how performance is affected in this context. Furthermore, the current study presented the faces on a body – which may be a more realistic depiction of the type of imagery that may be encountered in real environments.

1.4 The Present Study

The present study aims to understand how the use of morphed faces will impact on human face matching performance in a one-to-many face matching task. This study is an extension of the Robertson et al. (2017) study, where the current study is interested in replicating the results found in the 50% morph condition within a one-to-many face matching task. However, unlike that study, the morphs in the current study will have undergone image processing after the morphing procedure to remove any image artefacts (placing them onto donor heads and using Adobe® Photoshop® to finalise). Target imagery will include a variety of different morphing procedures, including ones where only 2 images, 8 images and 16 images (or constituent faces) were used to create the morphed image, and will also use non-morphed target images as controls.

Based on the face space literature (McIntyre et al., 2013), as well as previous research in this area (Robertson et al., 2017), the following hypotheses are presented:

H1 Participants will be most likely to declare a constituent face a match to its morph (commit a false alarm) when two images are used to create the morph, compared with other groups.

H2 As the number of faces used to create a morph increases, it will be harder to match a constituent face to its morph.

H3 If the same morphed face is placed on alternate images of the same donor it will pass as a match (predicting a null effect from the manipulation).

CHAPTER 2

Method

2.1 Ethics Statement

This study was approved by the Defence Science and Technology (DST) Group Human Research Ethics Committee (NSID 07/029) and the University of Adelaide Human Research Ethics Subcommittee (17/45). Participants were provided with a Participant Information Sheet (Appendix A) and provided their consent (Appendix B) before commencing the study.

2.2 Participants

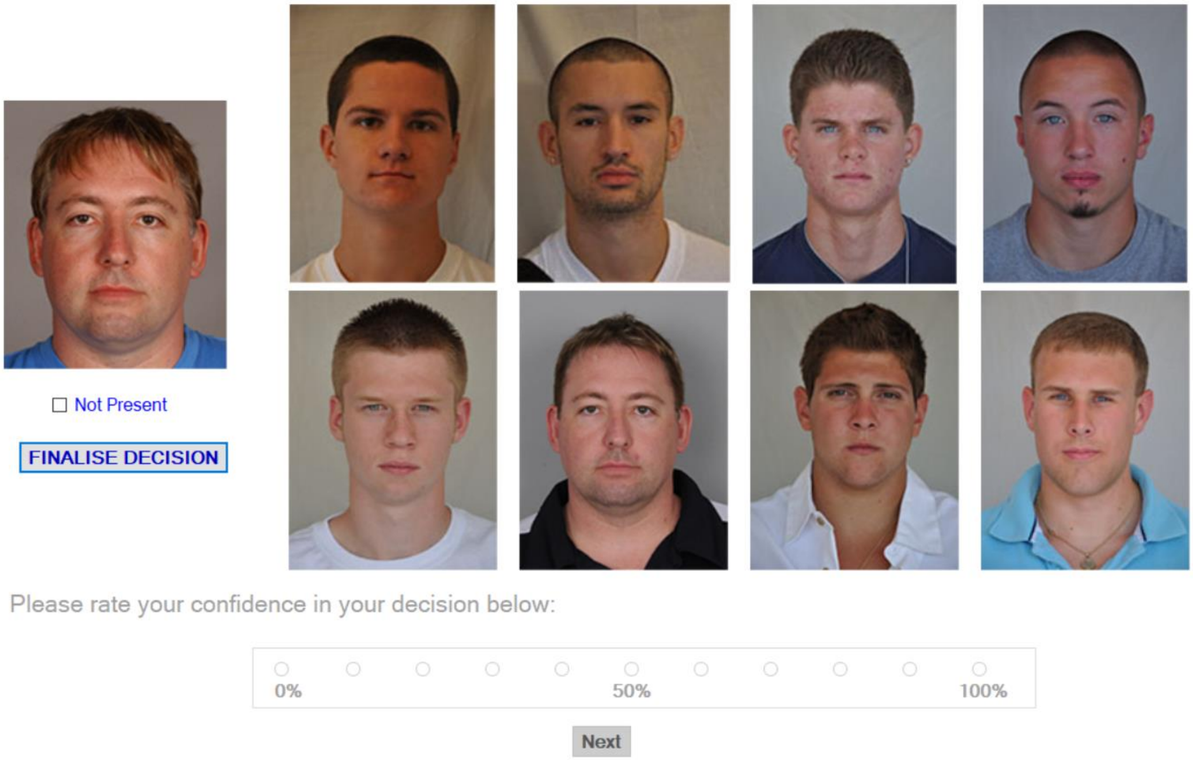
This study recruited 51 participants aged 18-67 years (25 females, mean age = 27.8, SD = 13.0). The majority of participants identified themselves as Caucasian (74.5 %), followed by Asian (23.5%) and mixed race (2.0%). Participants were recruited from the University of Adelaide first year psychology students (*N* = 7) and the general public (*N* = 44). The psychology students were recruited through the Research Participation System at the University of Adelaide and received course credit for participation. Participants from the general public, on the other hand, were recruited through posters placed around the University of Adelaide North Terrace campus (details of poster in Appendix C), and received no incentive for participation.

Participants were excluded from participation if they were younger than 18 years, were not proficient in English, and/or required vision correction (contact lenses or spectacles) but did not bring it to the study session.

2.3 Design

This one-to-many face matching study used a repeated measures design, with the number of images used to create the morphed image as the single manipulated variable. Each trial contained a candidate list with the exemplar image and 8 potential target images. A total of 4 conditions were formed based on the number of faces used to create the morphed image: Control (original faces that had no manipulation), 2-Image Morphs (used 2 images to create the morph), 8-Image Morphs (used 8 images) and 16-Image Morphs (used 16 images). Each candidate list was only seen once by each participant, and the order of the images in the candidate list, and the order in which the candidate lists were presented to participants, were randomised and counterbalanced for each participant.

In this study, the target was present in the trials with the manipulated faces when the morphed face in the exemplar image was also present in the candidate list, albeit presented on a different head (see Figure 1). The target was absent when the morphed face was not present in the candidate list, while the constituent faces (those used to make the morph) were included. In this scenario, the constituent faces acted as distractors, as the correct answer for these tasks required the participants to decide that the target was not present in the candidate list. On the other hand, for the control trials, the target was present when an alternate image of the target was also in the candidate list, while the target was absent when the candidate list did not include the target.



Please rate your confidence in your decision below:

Figure 1. Example of a trial where the target was present. The morphed face in the target image had also been superimposed onto one of the faces in the candidate list (specifically the second head from the left, on the bottom row).

The primary dependent variables were accuracy, confidence and response latency. For the purposes of calculating signal detection measures; a 'hit' was defined as correctly identifying the target image/morphed face in the candidate list when it was present, a 'miss' was defined as falsely identifying the target image/morphed face as not present in the candidate list when it was actually present, a 'correct rejection' was defined as not selecting an image in the candidate list, when the target image/morphed face was not present, and a 'false alarm' was defined as incorrectly selecting an image in the candidate list, if the target image/morphed face was not present.

2.4 Materials/Apparatus

2.4.1 Image source. Images used in this study were sourced from a range of facial databases collected for the purposes of research (Beveridge et al., 2014; X. Chen, Flynn, & Bowyer, 2003; Flynn, Bowyer, & Phillips, 2003; Phillips et al., 2011; Phillips, Moon, Rizvi, & Rauss, 2000; Phillips, Wechsler, Huang, & Rauss, 1998; Ricanek & Tesafaye, 2006).

2.4.1.1 Faces with manipulation: 2-Image morphs (Condition A). Twenty-eight pairs of faces were randomly selected from the facial databases. Gender was balanced, and these faces were not use elsewhere in the study. Each pair was matched for gender and ethnicity. Twenty-eight morphs were then created using the Fantamorph (http://www.fantamorph.com/index.html) program, by morphing the 28 pairs of images together one pair at a time. Post processing in Adobe® Photoshop® was used to place the morphed faces onto donor heads, and remove artefacts common in the morphing process (Robertson et al., 2017). Fourteen of the morphs were placed on two alternate donor heads, which were images of the same person taken at different times. A facial recognition algorithm supplied by DST Group was used to select the top eight matching images to each of the 28 morphs, resulting in 28 candidate lists. The target present candidate lists included the morph on the alternate donor head and the next seven top matching images, that did not include images used to make the morph (*n* = 14; see Figure 2). The target absent candidate list included the two images used to make the morph and the next six top matching images (*n* = 14; see Figure 3).

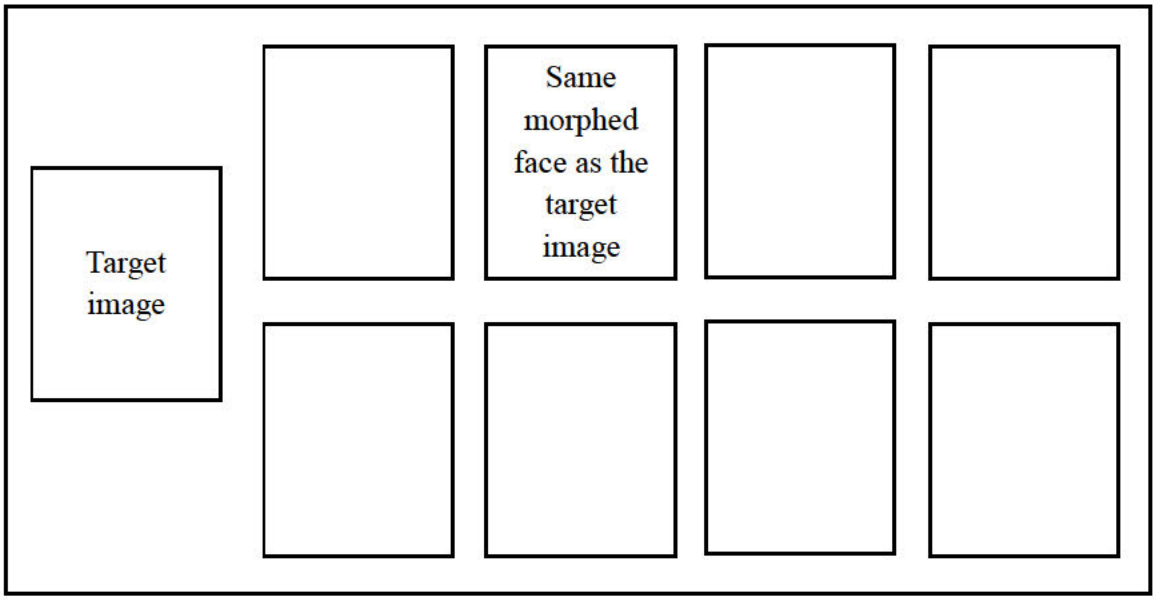


Figure 2. Example of how the trial screen was organised in the target present trials for the 2-Image morphs, 8-Image morphs and 16-Image morphs conditions. The location of the morphed image within the candidate list was randomised for each trial.

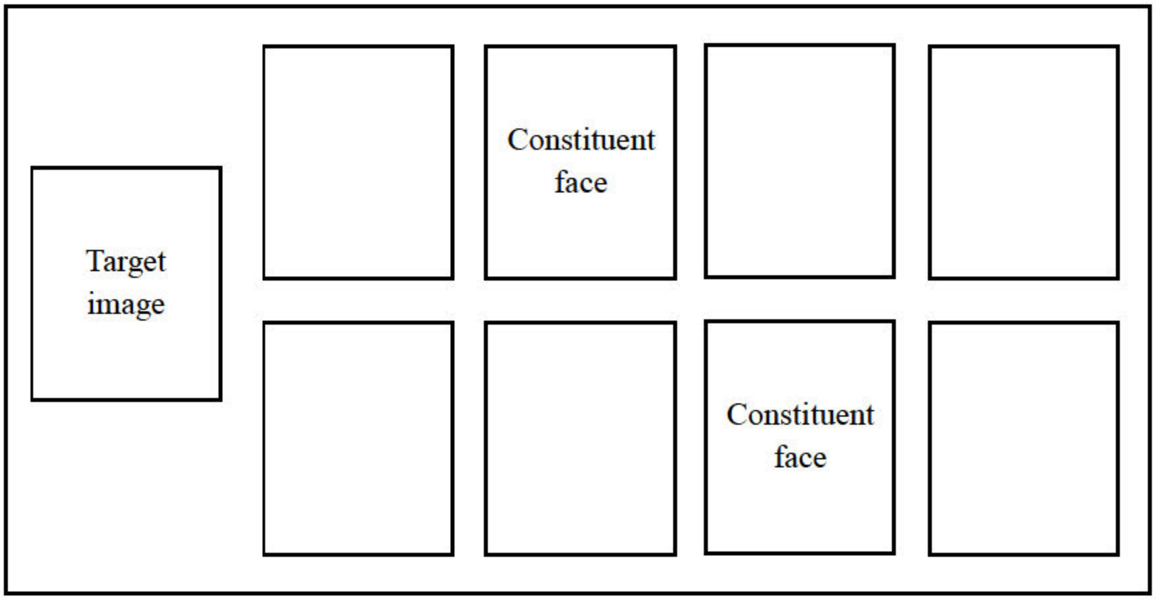


Figure 3. Example of how the trial screen was organised for the target absent trials, in the 2-Image morphs condition. The location of the constituent faces within the candidate list was randomised for each trial.

2.4.1.2 Faces with manipulation: 8-Image morphs (Condition B). Twenty-eight sets of faces were randomly selected from the facial databases. Gender was balanced, and these faces were not use elsewhere in the study. Each set of faces comprised of 8 faces, which were matched for gender and ethnicity. 28 morphs were then created using a Defence Science and Technology (DST) Group developed software tool, by morphing the 28 sets of eight images together one set at a time. Post processing in Adobe® Photoshop® was used to place the morphed faces onto donor heads, and remove artefacts common in the morphing process (Robertson et al., 2017). Fourteen of the morphs were placed on two alternate donor head, which were images of the same person taken at different times. A facial recognition algorithm supplied by DST Group was used to select the top eight matching images to each of the 28 morphs, resulting in 28 candidate lists. The target present candidate lists included the morph on the alternate donor head, and the next seven top matching images (*n* = 14; see Figure 2). The target absent candidate list included the eight images used to make the morph, which were also the top matching images (*n* = 14; see Figure 4).

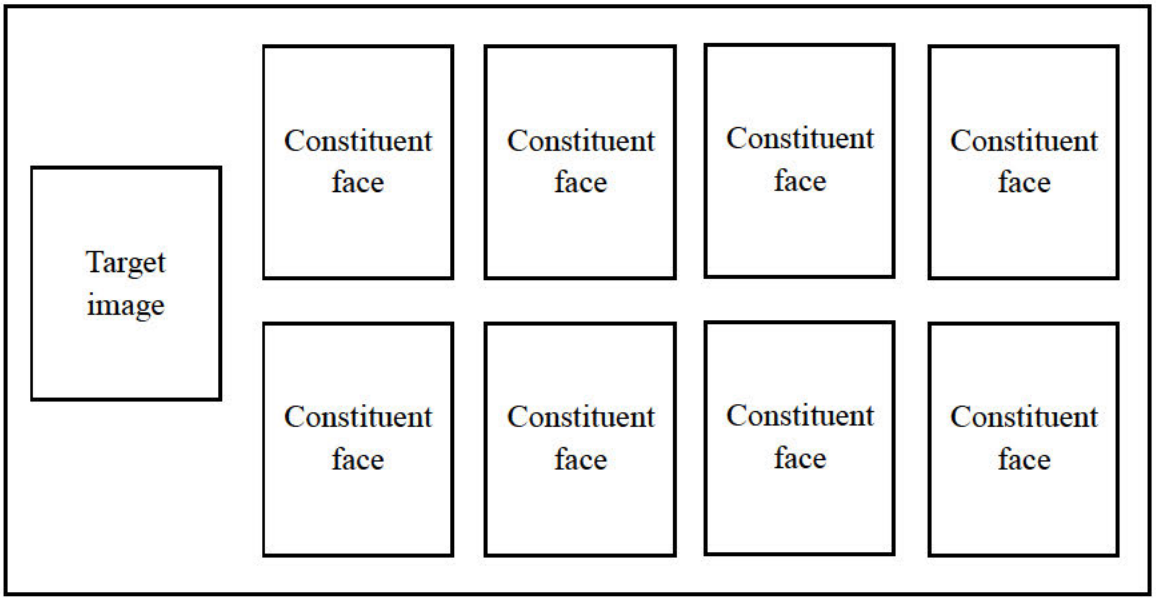


Figure 4. Example of how the trial screen was organised for the target absent trials, in the 8-Image morphs and 16-Image morphs conditions. The order of the constituent faces was randomised, so they were not presented in match score order.

2.4.1.3 Faces with manipulation: 16–image morphs (Condition C). Twenty-eight sets of faces were randomly selected from the facial databases. Gender was balanced, and these faces were not use elsewhere in the study. Each set of faces comprised of 16 faces, which were matched for gender and ethnicity. Twenty-eight morphs were then created using a DST Group developed software tool, by morphing the 28 sets of 16 images together one set at a time. Post processing in Adobe® Photoshop® was used to place the morphed faces onto donor heads, and remove artefacts common in the morphing process (Robertson et al., 2017). Fourteen of the morphs were placed on two alternate donor heads, which were images of the same person taken at different times. A facial recognition algorithm supplied by DST Group was used to select the top eight matching images to each of the 28 morphs, resulting in 28 candidate lists. The target present candidate lists included the morph on the alternate donor head, and the next seven top matching images, that had not been used to make the morph (*n* = 14; see Figure 2). The target absent candidate list included the top eight matching images that had been used to make the morph (*n* = 14, see Figure 4).

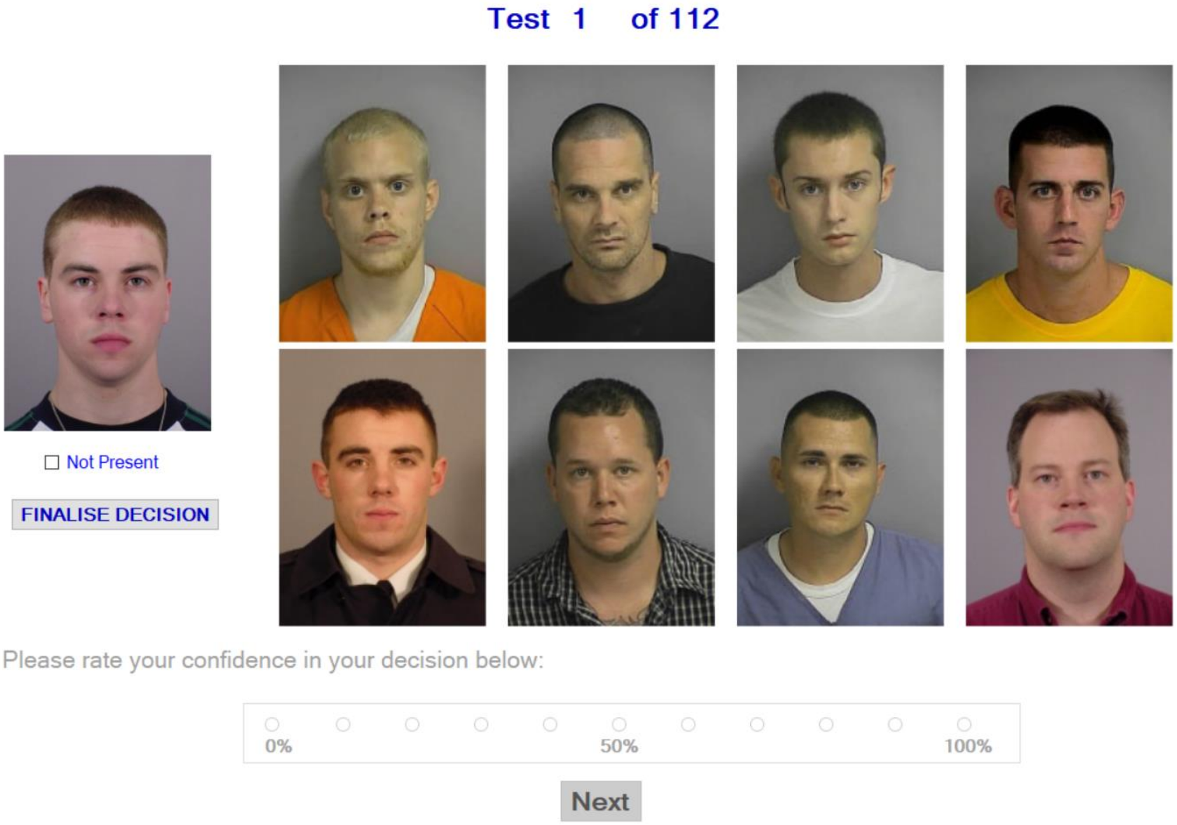
2.4.1.4 Faces with no manipulation: real faces (Condition D). Twenty-eight faces were randomly selected from the facial databases, and randomly allocated to two groups. Gender was balanced, and these faces were not used elsewhere in the study. A facial recognition algorithm supplied by DST Group was used to select the top eight matching images to each of the 28 faces from a database of 7890 images, resulting in 28 candidate lists. An alternate image of the target appeared in all 28 candidate lists, however it was removed from 14 of them and replaced with the next highest matching image to create the target absent candidate lists.

2.4.2 Duplicates. Once the imagery for the all the trials were selected, it was inspected for duplicates. If a duplicate was present, one of the images was removed and replaced with the next closest match.

2.4.3 Ordering of test sets. The number of trials was limited to 112 in order to reduce dropout rates. This study utilised a balanced design, with a 50:50 ratio of target present to target absent trials. Participants were randomly assigned to an order of presentation. All images present in the candidate list were randomised, and so were not presented in match score order – as may be expected if a face matching algorithm were used to select the faces in an operational setting.

2.5 Experimental Application

The experimental application was developed by a programmer employed by the DST Group. A screen shot of the application as shown to participants can be seen in Figure 5.



Please rate your confidence in your decision below:

Figure 5. Screenshot of a trial in the face matching application

2.6 Procedure

This study was conducted in the Applied Cognitive Experimental Psychology computer laboratory at the University of Adelaide (room 219 in the Hughes building). At the start of each session, each participant was given a verbal briefing of the study, along with a copy of the participant information sheet (Appendix A). Following this, participants gave their consent by clicking on the designated button on the first screen of the experimental interface (Appendix B). At this point, each participant was allocated a unique identification number by the experimental interface to facilitate data management. The following screen then collected information about the participants' demographics, including age, gender and ethnicity, as well as information about their vision.

The next screen contained specific instructions about the face matching task, which were also read by the researcher (Appendix D). After that, participants were required to perform two practice tasks (Appendix E). Performance feedback was given for those trials. Participants were also reminded that they were not looking for identical images, and advised that there will be some form of differences between images of the same person. More importantly, participants were not informed of the exact aims of the experiment. This was to avoid priming the participants, as it may have distracted them with a need to be on the lookout for manipulated faces, potentially impacting on response time and confidence in particular (Duncan, 2006).

The next succession of screens were the self-paced experimental trials, with an identical interface to the practice trials. During each trial, participants were presented with a target image and a candidate list of eight images – displayed side-by-side, and were required to decide if the target was present in the candidate list. If the target was present, participants had to click on the image, if not, they were required to click the 'not present' button. After that, participants were asked to rate their confidence in their decision, on a 0% to 100% scale with 10% increments. At the same time, the application recorded the accuracy, confidence and response latency (recorded as the time elapsed between the display of the images and when the participant finalised their decision) of each face matching decision. Finally, at the end of the experiment, participants had the option of providing their email address to receive overall feedback on their performance.

CHAPTER 3

Results

3.1 Data Screening, Assumptions and Test Selection

Data were screened prior to analysis to check for missing data and to assess normality. No missing data were found. Half of the variables had a Shapiro-Wilk statistic result of $p < .001$, and some of them were heavily skewed (skewness ranged from -3.21 to 3.08; see Appendix F), so the dataset did not meet assumptions of normality. However, data transformation was not undertaken, as the skewness of some of the variables, like confidence, accuracy and response latency, are valid representations of the population distribution (Burton, White, & McNeill, 2010). Hence, non-parametric tests, which were appropriate for within-subject designs, were used instead. The Wilcoxon sign-rank test was used to assess differences between two groups, while the Friedman ANOVA by ranks rest was used to assess differences between more than two groups. Alpha was set as $\alpha = .05$ with two-tailed significance values for the Friedman ANOVA by ranks tests, while a Bonferroni correction was also applied for all post-hoc Wilcoxon signed-rank tests. Effect size for the Wilcoxon sign-rank test was calculated using the formula $r = Z/\sqrt{N}$, and all effect size interpretations were made with reference to Cohen (1988).

3.2 Overall performance

Tables 1-3 shows the descriptive statistics for accuracy, confidence and response latency across all experimental groups. The responses for both the ‘target present’ and ‘target absent’ conditions were combined to form an ‘overall’ score. The following analyses were conducted across the difference between those experimental groups, in respect to the overall scores. A Bonferroni correction was applied such that all post hoc Wilcoxon signed-rank tests were reported at a .008 level of significance (two-tailed).

Table 1

Descriptive Statistics for Accuracy (%) by Experimental Groups

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	88.25	95.96	92.86	87.90
Median	93.00	96.00	93.00	89.00
SD	10.68	5.57	7.24	10.12
Variance	114.15	31.00	52.48	102.49
Minimum	61.00	71.00	61.00	64.00
Maximum	100.00	100.00	100.00	100.00

Note. $N = 51$ for all groups.

3.2.1 Accuracy. Descriptive statistics suggested accuracy was highest in the 8-Image Morphs group, followed by the 16-Image group, 2-Image group and Control group (see Table 1). A Friedman ANOVA by ranks test indicated that accuracy was significantly different across all four groups ($\chi^2(3) = 46.96$, $p < .001$, $W = .31$). Wilcoxon signed-rank tests revealed a large and significant difference between the 2-Image Morphs and 8-Image Morphs groups ($Z = -4.93$, $p < .001$, $r = -.69$), the 2-Image Morphs and 16-Image Morphs groups ($Z = -3.48$, $p < .001$, $r = -.49$), the 8-Image Morphs and 16-Images Morphs groups ($Z = -3.54$, $p < .001$, $r = -.50$), the 8 images and control groups ($Z = -5.31$, $p < .001$, $r = -.74$), and the 16-Image Morphs and Control groups ($Z = -3.67$, $p < .001$, $r = -.51$). However, no significant difference was found between the 2-Image Morphs and Control groups ($Z = -0.32$, $p = .75$, $r = -.045$).

Therefore, overall accuracy was the highest when 8 constituent faces were used to create the morph, followed by 16 faces and 2 faces, and the lowest when the control faces (no morphs) were used.

Table 2

Descriptive Statistics for Confidence (%) by Experimental Group

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	79.45	83.13	80.14	79.75
Median	80.71	83.57	78.21	80.00
SD	11.00	11.53	10.91	11.51
Variance	121.05	133.05	118.94	132.57
Minimum	45.00	45.71	47.86	38.93
Maximum	100.00	100.00	98.93	98.21

Note. $N = 51$ for all groups.

3.2.2 Confidence. Descriptive statistics suggested confidence was the greatest when participants were exposed to the 8-Image Morphs group, followed by the 16-Images group, 2-Images group, and Control group (see Table 2). A Friedman ANOVA by ranks tests indicated that confidence was significantly different across the four groups ($\chi^2(3) = 46.52$, $p < .001$, $W = .30$). Wilcoxon signed-rank tests revealed a large and significant difference in confidence levels between the 2-Image Morphs and 8-Image Morphs groups ($Z = -5.47$, $p < .001$, $r = -.77$), the 8-Image Morphs and 16-Image Morphs groups ($Z = -5.17$, $p < .001$, $r = -.72$), and the 8-Image Morphs and Control groups ($Z = -4.66$, $p < .001$, $r = -.65$). No significant differences were found between the 2-Image Morphs and 16-Image Morphs groups ($Z = -1.89$, $p = .058$, $r = -.27$), and the 8-Image Morphs and Control groups ($Z = -1.08$, $p = .28$, $r = .15$), however, they had a medium and small effect size respectively. Additionally, the 2-Image Morphs and Control groups had no significant differences and a negligible effect size ($Z = -0.52$, $p = .60$, $r = -.073$).

Therefore, overall confidence was the highest when 8 faces were used to create the morphed face, followed by 16 faces and control faces (no morphs), and the lowest when only 2 faces were used to create the morph.

Table 3

Descriptive Statistics for Response Latency (s) by Experimental Group

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	13.26	10.16	11.63	14.35
Median	12.27	9.83	11.33	13.73
SD	6.00	3.89	4.85	5.81
Variance	35.96	15.12	23.54	33.76
Minimum	4.78	3.89	4.26	4.55
Maximum	34.60	19.07	25.29	28.33

Note. $N = 51$ for all groups.

3.2.3 Response latency. Descriptive statistics suggested response latency was fastest in the 8-Image Morphs group, followed by the 16-Images, 2-Images and Control groups (see Table 3). A Friedman ANOVA by ranks test indicated that response latency was significantly different across all four groups ($\chi^2(3) = 86.53$, $p < .001$, $W = .57$). Wilcoxon signed-rank tests revealed a large and significant difference in response latency between the 2-Image Morphs and 8-Image Morphs groups ($Z = -5.62$, $p < .001$, $r = -.70$), the 2-Image Morphs and 16-Image Morphs groups ($Z = -4.50$, $p < .001$, $r = -.63$), the 8-Image Morphs and 16-Image Morphs groups ($Z = -4.30$, $p < .001$, $r = -.84$), the 8-Image Morphs and Control groups ($Z = -5.98$, $p < .001$, $r = -.84$), and the 16-Image Morphs and Control groups ($Z = -5.61$, $p < .001$, $r = -.79$). There was also a moderate difference in response latency between the 2-Image Morphs and Control groups, but the difference was not significant ($Z = -2.58$, $p = .01$, $r = -.36$).

Therefore, response latency was the fastest when 8 constituent faces were used to create the morphed face, followed by 16 faces and 2 faces, and the slowest in the control group (no morphs).

3.3 Hypothesis 1: Participants will be most likely to declare a constituent face a match to its morph (commit a false alarm) when two images are used to create the morph, compared with other groups

To investigate this hypothesis, it was necessary to determine if false alarms were highest when a simple morphed face (using 2 images) was presented. To achieve this, false alarm rate for the 2-Image Morphs group and the three other experimental groups was compared. Table 4 shows the descriptive statistics for false alarm across the groups. A Bonferroni correction was applied such that all post hoc Wilcoxon signed-rank tests were reported at a .017 level of significance (two-tailed).

Table 4

Descriptive Statistics for False Alarm by Experimental Groups

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	.20	.05	.07	.16
Median	.14	.00	.00	.14
SD	.20	.09	.13	.17
Variance	.04	.008	.02	.03
Minimum	.00	.00	.00	.00
Maximum	.71	.50	.71	.57

Note. $N = 51$ for all groups.

The hypothesis predicted that the rate of false alarms would be higher in the 2-Image Morphs group. Descriptive statistics suggested that the false alarm rate was the highest in the 2-Image Morphs group, compared to the Control, 16-Images and 8ontrol groups (see Table 3). Wilcoxon signed-rank tests revealed a large and significant difference in false alarm rates between the 2-Image Morphs and 8-Image Morphs groups ($Z = -4.82$, $p < .001$, $r = -.68$), and the 2-Image Morphs and 16-Image Morphs groups ($Z = -4.49$, $p < .001$, $r = -.63$). There was also a moderate but non-significant difference between the 2-Image Morphs and Control groups ($Z = -1.88$, $p = .06$, $r = -.26$).

Therefore, these results did not support the hypothesis. While the false alarm rate was significantly higher when the participant was presented with morphed images made from 2 faces, compared to the morphed images made from 8 or 16 images, there was no significant difference compared to when the participant was presented with the control faces (no morph).

3.4 Hypothesis 2: As the number of faces used to create a morph increases, it will be harder to match a constituent face to its morph

To investigate this hypothesis, it was necessary to determine if face matching performance improves (false alarm and response latency decreases, and confidence increases) when the number of faces used to create the morph increases. Measures of false alarm, confidence and response latency, within the target absent condition, were compared across the 2-Image Morphs, the 8-Image Morphs, and the 16-Image Morphs groups. A Bonferroni correction was applied, so that all post-hoc Wilcoxon signed-rank tests were reported at a .017 level of significance (two-tailed). Table 5 shows the descriptive statistics for false alarm, confidence and response latency across the three groups.

Table 5

Descriptive Statistics for False Alarm, Confidence and Response Latency by Experimental Group, Within the 'Target Absent' Condition

	2-Image Morphs			8-Image Morphs			16-Image Morphs		
Statistics	FA (0-1)	C (%)	RL (s)	FA (0-1)	C (%)	RL (s)	FA (0-1)	C (%)	RL (s)
M	.20	75.73	18.67	.05	77.48	14.04	.07	76.20	14.66
Median	.14	75.71	16.80	.00	76.43	13.65	.00	75.71	14.48
SD	.20	17.59	9.28	.09	14.09	5.69	.13	13.62	6.62
Variance	.04	309.55	86.21	.008	198.60	32.37	.02	185.38	43.86
Min.	.00	0.93	5.34	.00	34.29	4.22	.00	40.71	4.94
Max.	.71	100.00	57.41	.50	100.00	28.04	.71	99.29	33.60

Note. *N* = 51 for all groups. FA = false alarm, C = confidence, RL = response latency.

3.4.1 False alarm. The hypothesis predicted that the false alarm rate will be the lowest in the 16-Image Morphs group, followed by the 8-Image and the 2-Image group. However, descriptive statistics suggested false alarm rates were the lowest when participants were exposed to images from the 8-Image Morphs group, followed by the 16-Images group and the 2-Images group (see Table 5). A Friedman ANOVA by ranks test indicated that the false alarm rates were significantly different across the three groups ($\chi^2(2) = 40.98$, $p < .001$, $W = .40$). Wilcoxon signed-rank tests revealed a large and significant difference in the false alarm rates between the 2-Image Morphs and 8-Images groups ($Z = -4.82$, $p < .001$, $r = -.68$), and the 2-Images and 16-Images groups ($Z = -4.49$, $p < .001$, $r = -.63$). There was also a small but non-significant difference between the 8-Images and 16-Images groups ($Z = -1.26$, $p = .21$, $r = .18$).

Therefore, these results did not support the hypothesis. The false alarm rate was not significantly different when the participant was presented with morphed images made from 8 faces or 16 faces, although the higher false alarm rate with the morphed images made using 2 faces was anticipated.

3.4.2 Confidence. The hypothesis predicted that confidence will be the highest in the 16-Images group, followed by the 8-Images group and the 2-Images group. However, descriptive statistics suggested that confidence was the highest when participants were exposed to images from the 8-Images group, followed by the 16-Images group and the 2-Images group (see Table 5). A Friedman ANOVA by ranks test indicated that there was no significant difference in confidence across the three groups ($\chi^2(2) = 5.65$, $p = .06$, $W = .06$). Wilcoxon signed-rank tests revealed a moderate but non-significant difference in confidence between the 8-Images and 16-Images groups ($Z = -2.05$, $p = .04$, $r = .29$). A negligible and non-significant difference also existed between the 2-Images and 8-Images groups ($Z = -0.19$, $p = .85$, $r = -.03$), and the 2-Images and 16-Images groups ($Z = -0.19$, $p = .85$, $r = -.03$).

Therefore, the results did not support the hypothesis, as there was no significant difference in accuracy when participants were presented with morphed faces made from either 2 faces, 8 faces or 16 faces.

3.4.3 Response Latency. The hypothesis predicted that response latency will be the fastest in the 16-Images group, followed by the 8-Images group and the 2-Images group. However, descriptive statistics suggested response latency was the fastest when participants were exposed to images from the 8-Images group, followed by the 16-Images group and the 2-Images group (see Table 5). A Friedman ANOVA by ranks test indicated that response latency was significantly different across the three groups ($\chi^2(2) = 34.16$, $p < .001$, $W = .34$). Wilcoxon signed-rank tests revealed a large and significant difference in response latency between the 2-Images and the 8-Images groups ($Z = -5.40$, $p < .001$, $r = -.76$), and the 2-Images and 16-Images groups ($Z = -5.31$, $p < .001$, $r = -.74$). There was also a small but non-significant difference between the 8-Images and 16-Images group ($Z = -1.11$, $p = .27$, $r = -.16$).

Therefore, the results did not support the hypothesis, as response latency was not significantly different when the participants were presented with morphed images made from 8 faces or 16 faces, although the slower response latency with the morphed images made using 2 faces was anticipated.

3.5 Hypothesis 3: If the same morphed face is placed on alternate images of the same donor it will pass as a match (predicting a null effect from the manipulation)

In order to investigate this hypothesis, it was necessary to determine if face matching performance remains the same in the target present condition, regardless of the number of faces used to create the morph or whether control faces were used. In other words, the manipulation here was hypothesised to have a null effect. Measures of hit rate, confidence and response latency (target present condition) were compared across the experimental groups. A Bonferroni correction was applied, so that all post-hoc Wilcoxon signed-rank tests were reported at a .008 level of significance (two-tailed). Tables 6-8 shows the descriptive statistics for hit rate, confidence and response latency across the groups.

Table 6

Descriptive Statistics for Hit Rate by Experimental Groups, Within the 'Target Present' Condition

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	.96	.98	.93	.92
Median	1.00	1.00	.93	.93
SD	.07	.05	.06	.08
Variance	.005	.003	.004	.007
Minimum	.64	.71	.79	.64
Maximum	1.00	1.00	1.00	1.00

Note. $N = 51$ for all groups.

3.5.1 Hit. The hypothesis predicted that the rate of hits will be equal across all four groups. However, descriptive statistics suggested hit rate was the highest when participants were exposed to images from the 8-Image Morphs group, followed by the 2-Images group, the 16-Images group and the Control group (see Table 6). A Friedman ANOVA by ranks test indicated that there was a significant difference in hit rate across the four groups ($\chi^2(3) = 29.60, p < .001, W = .19$). Wilcoxon signed-rank tests revealed a large and significant difference in hit rate between the 8-Images and 16-Images groups ($Z = -3.76, p < .001, r = -.53$), and the 8-Images and Control groups ($Z = -4.00$, $p < .001$, $r = -.56$), and a moderate and significant difference between the 2-Images and Control groups ($Z = -3.12$, $p = .002$, $r = -.44$). There was also a moderate but non-significant difference between the 2-Images and 16-Images groups ($Z = -2.63$, $p = .009$, $r = -.37$), and a small but non-significant difference between the 2-Images and 8-Images groups ($Z = -1.75$, $p = .08$, $r = -.25$), and the 16-Images and Control groups ($Z = -0.90$, $p = .37$, $r = -.13$). This did not support the predicted null hypothesis, as there was a significant difference in hit rate across groups.

Table 7

Descriptive Statistics for Confidence (%) by Experimental Groups, Within the 'Target Present' Condition

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	86.11	88.78	84.08	85.55
Median	88.57	91.43	85.00	86.43
SD	11.34	10.82	10.01	10.23
Variance	128.63	117.04	100.15	104.73
Minimum	52.14	57.14	55.00	50.71
Maximum	100.00	100.00	99.29	100.00

Note. $N = 51$ for all groups.

3.5.2 Confidence. The hypothesis predicted that confidence will be equal across all four groups. Confidence was the highest when participants were exposed to images from the 8-Image Morphs group, followed by the 2-Images group, the Control group and the 16-Images group (see Table 7). A Friedman ANOVA by ranks test indicated that there was a significant difference in confidence across the four groups ($\chi^2(3) = 43.23$, $p < .001$, $W = .28$). Wilcoxon signed-rank tests revealed a large and significant difference in confidence between the 2-Images and 8-Images groups ($Z = -3.54$, $p < .001$, $r = -.50$), the 8-Images and 16-Images groups ($Z = -5.66$, $p < .001$, $r = -.79$), and the 8-Images and Control groups ($Z = -3.94$, $p < .001$, $r = -.55$). Additionally, there was a moderate and significant difference between the 2-Images and 16-Images groups ($Z = -2.94$, $p = .003$, $r = -.41$). There was also a small but non-significant difference between the 2-Images and Control groups ($Z = -1.61$, $p = .11$, $r = -.23$), and the 16-Images and Control groups ($Z = -1.70$, $p = .09$, $r = -.24$). This did not support the predicted null hypothesis, as there was a significant difference in confidence across groups.

Table 8

Descriptive Statistics for Response Latency (s) by Experimental Groups, Within the 'Target Present' Condition

Statistic	2-Image Morphs	8-Image Morphs	16-Image Morphs	Control
M	7.85	6.28	8.60	10.19
Median	7.01	5.90	7.24	9.50
SD	4.10	2.84	4.03	4.26
Variance	16.83	8.06	16.24	18.18
Minimum	2.66	2.36	3.15	3.46
Maximum	25.22	13.65	21.80	24.28

Note. $N = 51$ for all groups.

3.5.2 Response latency. The hypothesis predicted that the response latency will be equal across all four groups. Response latency was the fastest when participants were exposed to images from the 8-Image Morphs group, followed by the 2-Images group, the 16-Images group and the Control group (see Table 8). A Friedman ANOVA by ranks test indicated that there was a significant difference in response latency across the four groups ($\chi^2(3) = 71.47, p < .001, W = .47$). Wilcoxon signed-rank tests revealed a large and significant difference in response latency between the 2-Images and 8-Images groups ($Z = -4.77, p < .001, r = -.67$), the 2-Images and Control groups ($Z = -4.58, p < .001, r = -.64$), the 8-Images and 16-Images groups ($Z = -5.76, p < .001, r = -.81$), the 8-Images and Control groups ($Z = -6.14$, $p < .001$, $r = -.86$), and the 16-Images and Control groups ($Z = -3.62, p < .001, r = -.51$). There was also a small and significant difference between the 2-Images and 16-Images groups ($Z = -2.10$, $p = .04$, $r = -.29$). This did not support the predicted null hypothesis, as there was a significant difference in response latency across groups.

CHAPTER 4

Discussion

This study aimed to understand how the usage of morphed images impacted human face matching performance in a simultaneous one-to-many face matching task. The results were mixed in support of the hypotheses. This discussion will focus on the methodological, theoretical and applied implications of the study.

4.1 Overall Performance

Face matching accuracy and confidence was greatest and response latency was the fastest, overall, with the morphed faces made from 8 images, followed by the morphed faces made from 16 images. Performance was the worse when 2 images were used to create the morph, and in the no morphs control group. The pattern of results found here were somewhat surprising. However, this can be explained by appealing to the mechanisms of face space, which will be discussed later.

Moreover, the overall accuracy for all four groups was also very high (ranging from 87.90% to 95.96%), compared to the landmark study by Bruce et al. (1999). This could be due to the quality of the images used, as the target images in Bruce et al. (1999) were of comparably lower quality, compared to the images used within the current study. Therefore, the images used here would contain more featural information, which makes this face matching task much easier.

4.2 Hypothesis 1: Participants will be most likely to declare a constituent face a match to its morph (commit a false alarm) when two images are used to create the morph, compared with other groups

There was no significant difference in the false alarm rate when participants were presented with morphed faces made from 2 images, compared to the control (no morphs) group, which did not support the hypothesis. The false alarm rates within the control condition were consistent with previous studies that looked at unfamiliar face matching in target absent trials, indicating that this is not a difficult face matching task (Bruce et al., 1999; Megreya & Burton, 2006, 2007). Therefore, this pattern of results suggests that, contrary to Robertson et al. (2017), the usage of 2-Image morphs is not an effective form of evading detection and allowing the false acceptance of an image. This difference in results may be caused by two factors. Firstly, compared to the current study, Robertson et al. (2017) used more difficult face pairs, as participants had to compare faces of different sizes and with different backgrounds. This can increase the rate of false alarms. Secondly, the current study morphed together two random images to create the 2-Image morphs. As seen in Figure 6, this means that the both constituent images could be located anywhere in face space, from being near to each other or far apart, with the morphed image lying on the midpoint of the vector that joins the two constituent images (Busey, 1998). When the morph is presented as the target image, it samples nearby items in participants' face space, which allows them to make a judgement on similarity (Busey & Tunnicliff, 1999). If any images in the candidate list are located within that nearby region in face space, the participant may falsely select that image as a match. However, if the constituent images were located in far apart from each other in face space, they may not fall under that sampling region, thus they will not be erroneously chosen as a match. Therefore, the inclusion of the constituent faces in the candidate list may have had a varying effect on the false alarm rate, and this is supported by the relatively high variance in the 2-Image morphs group.

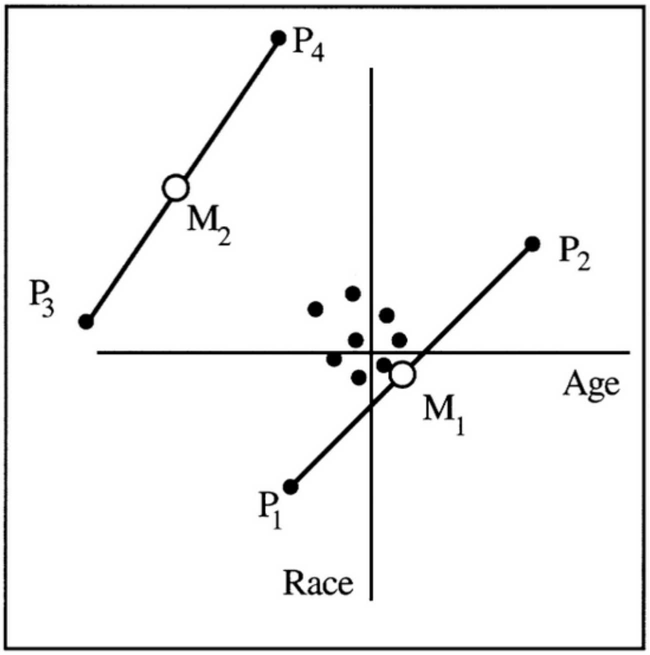


Figure 6. Illustration of how the location of parent faces (P_1 to P_4) within face space influences the location of the morphed faces (M_1 and M_2; Busey, 1998).

More interestingly, the usage of 8-Image morphs and 16-Image morphs resulted in a lower false alarm rate. Again, this can be explained using psychological face space. Morphed faces are usually located closer to the centre of face space (Busey, 1998). However, Busey (1998) also found that morphs which are predicted to be located nearer to the centre of face space, which would be the case for morphs made from more than two constituent faces, tend to shift away from the centre, making them less typical than predicted. Therefore, this could have enhanced their discriminability, leading to the much lower false alarm rates found in this study.

4.3 Hypothesis 2: As the number of faces used to create a morph increases, it will be harder to match a constituent face to its morph

The hit rate was the lowest, and response latency was the slowest, with morphs created using 2 constituent images, followed by those created using 8 or 16 images. There was also no significant difference between all three groups, when comparing their confidence levels. However, lack of a significant difference in confidence levels has a minimal impact on the overall results, as previous studies have shown that confidence and accuracy are not related in face matching trials where the target was absent (Stephens, Semmler, & Sauer, 2017).

The reason why 2-Image morphs have higher false alarm rates and lower response latency is due to how the candidate list was populated and the location of the morphs in psychological face space. For the 2-Image morphs condition, the candidate list was populated with the two constituent faces, and the rest was filled with the six highest matching non-constituent faces. On the other hand, the 8-Image morphs and the 16-Image morphs conditions were populated with the highest matching constituent images only. This created a confound, as it is unknown whether the higher false alarm rates could be attributed to the lower number of parent faces used, or the inclusion of the non-constituent faces in the candidate list. As the non-constituent faces were the highest matching faces to the morphed face, they could be located closer in faces space to the morphed face, compared to the parent image (Busey, 1998). Therefore, this could have erroneously increased the false alarm rates in the 2-Image morphs condition.

Moreover, the 8-Image morphs and 16-Image morphs groups had no significant difference in false alarm rate or response latency, which did not support the hypothesis. This is due to the location of the morphs in face space. As stated above, with a morph made from two faces, the morph is located at the midpoint of the vector that connects both constituent faces (Busey, 1998).

The same would apply to morphs made from more than two constituent faces, but in this scenario, the morph is hypothesised to be located in a location which equalises the difference between the morph and the constituent faces. As the constituent faces were chose by random, and thus can be quite dissimilar, they will be scattered randomly throughout face space. Therefore, the spatial distance between the morph and its constituent faces, on average, will be similar for both the 8-Image morphs and the 16-Image morphs, resulting in the similar false alarm rates and response latency.

4.4 Hypothesis 3: If the same morphed face is placed on alternate images of the same donor it will pass as a match (predicting a null effect from the manipulation)

The hit rate and confidence levels were the highest, and response latency was the fastest, with morphs made from 8 constituent images. Participants also performed better when the 2-Image morphs and 16-Image morphs were used, although this difference was not significant. Therefore, this did not support the hypothesis, as there was a significant difference between the 8-Image morphs group and the other three groups within all three performance variables (2-Image morphs, 16-Image morphs and control groups). This suggests that contrary to previous understandings, the usage of 8-Image morphs actually increased human face matching performance when the target image was also present in the candidate list.

This pattern of results may be explained by the uncanny valley effect, which was not accounted for during the construction of the hypotheses. This effect proposes that people find hyper-photorealistic computer-generated faces unsettling, due to their non-human imperfections and potential abnormal features (MacDorman, Green, Ho, & Koch, 2009; Seyama & Nagayama, 2007). Previous studies have shown that it is easier to discriminate between human-looking faces and fake-looking faces (Cheetham, Suter, & Jäncke, 2011). Therefore, as the matching image for each morphed target image is the only non-genuine image within the candidate list, this make it easier to select the corresponding matching face, resulting in the higher performance. Moreover, as the number of faces used to make the morph increases, the uncanny valley effect is expected to be amplified, as the resultant morphed faces will have a more photorealistic texture and potentially distorted facial proportions (MacDorman et al., 2009). This would contribute to the higher performance in the 8-Images morph group, compared to the 2-Images morph group. However, after a certain threshold, the additional faces used to create the morph face would eventually balance out the unnatural features, making the resultant morph face seem more natural (Seyama & Nagayama, 2007). This could reduce the discriminability of the morphed face, potentially explaining the lower performance of the 16-Images morphs, as compared to the 8-Images morphs.

Additionally, the way in which the matching image for the morphed target images was created could also have contributed to the higher performance in the 2-Image morphs, 8-Image morphs and 16-Image morphs groups, as compared to the control group. The matching images were created by placing the morphed face on alternate images of the same donor head that was used in the target image. The idea behind this process is to mimic the presentation of a candidate list in an operational setting, which would include the external features of a face. This is important, as people tend to reply more on external features when matching unfamiliar faces (Butavicius, Lee, & Vast, 2006). Also, this process would have introduced slight variations in pose, expression or lighting for the faces themselves. However, the morphing process could have made those differences insignificant. So, compared to the matching image for the control face, whilst both the matching images for the control and morphed faces had differences in the external features, the ones used for the morphed faces were much more similar in terms of the face itself. This would have resulted in better face matching performance when the participants were presented with morphed faces, and this was reflected in the results.

4.5 Study Implications

Overall, this study found that the usage of dissimilar morphs does not adversely impact human face matching performance, and in some scenarios, can even enhance performance. Strangely enough, the results found here contradicted a similar study conducted on facial recognition algorithms, which concluded that such algorithms are vulnerable to morphing image alterations (Ferrara et al., 2014). This further highlights the difference in how humans and facial recognition algorithms process faces. Previous studies have found that humans perform better when matching difficult face matching pairs, which includes morphs (Phillips & O'Toole, 2014). Therefore, it is dangerous to rely just on facial recognition algorithms to perform face matching tasks, and that it is vital to continue relying on human input for such tasks.

While morphs were found to have a minimal impact on human face matching performance, this study was conducted within an experimental setting, so there may be a different impact within an operational setting. Therefore, the knowledge gained here can be used to minimise any potential impact from morphs. Anyone who works in a job that relies on facial identification should be trained to recognise morphs, with feedback provided on their performance to maximise learning (White, Kemp, Jenkins, & Burton, 2014).

Additionally, certain policy changes are recommended to reduce any potential usage of morphs in identity documents. Currently, in Australia, applicants are able to hand in their own colour photos while applying for a passport or other identity documents (Department of Foreign Affairs and Trade, 2017). Therefore, this system can be potentially abused if the applicant uses a morphed face as their application image. And with the introduction of automated face recognition programs for immigration (i.e. SmartGate) in certain airports in Australia, which removes the human component in identity checks during immigration or emigration, the potential for abuse becomes even more worrying, as such automated systems are especially vulnerable to identity fraud via morphed faces (Ferrara et al., 2014). Therefore, the individuals involved in the verification of the applicants' identity should have additional training in the discrimination and recognition of morphed faces. Alternatively, the authorities involved could request the photograph to be taken when the passport application was lodged, to reduce the chances of the authentic photograph being replaced by a morphed image.

4.6 Strengths

A strength of this study was the usage of both the 'target present' and 'target absent' conditions within the study, which was the main limitation of previous research (e.g. Robertson et al., 2017). Additionally, this study also considered differences in confidence levels and response latency, which were not included in previous studies (e.g. Robertson et al., 2017).

4.7 Limitations

One major limitation was the ethnicity of the faces used in this study, which were sourced from predominantly Caucasian databases. This makes the findings here less applicable to face matching procedures in ethnically diverse countries, like Australia, where both the image subjects and facial reviewers may be of various ethnicities. Furthermore, while the current study had a significant proportion of Caucasian participants (74.5%), it did not consider the effect from the own-race bias, which suggests that people perform better on face matching tasks which contain faces consistent with their own race (Meissner & Brigham, 2001). This could have increased performance throughout the study, resulting in higher accuracy and confidence, as well as faster response latency.

Additionally, as stated above, there could be a possible confound in the study design. Within the 'target absent' condition, the candidate list for the morphs made from 2 constituent images was populated with both the two constituent images and the top six matching unrelated faces. This is different from the candidate list for the 8-Images morphs and 16-Images morphs groups, which only used all eight of the constituent faces or the top eight matching constituent faces respectively. Therefore, with the difference in the candidate lists, it is unknown whether the difference in performance between the 2-Images morphs and the other two groups could be attributed to the inherent difference in the morphed faces, or whether it came about due to the inclusion of the unrelated faces. This reduces the validity of the findings in the first and second hypotheses.

4.8 Suggestions for Future Research

Due to the inherent confounds present within the study design, the current study needs to be replicated in order to ensure the validity of the findings. To prevent further confounds, future studies should have a one-to-one study design, with the two highest matching faces for each morph condition used as the comparison image for non-mated trials. This should enable equal comparisons between groups, allowing for stronger conclusions to be drawn from the study results.

Additionally, this study was conducted with morphs made from random faces, implying that the majority of morphs were made by combining dissimilar faces. In psychological face space, similar faces would be located closer together, compared to dissimilar faces. The morphs made from similar faces would logically be more similar to their constituent faces, compared to those created using dissimilar faces. Therefore, future studies can also compare the performance of similar and dissimilar morphs, to see if the usage of similar morphs can pose an even greater security risk.

Another possibility is to get participants to rate the similarity of the 2-Image morphs, 8-Image morphs, 16-image morphs to their constituent faces, to produce a similarity rating. The process enables the faces to be mapped onto a face space model using multidimensional scaling, allowing for the calculation of the relative distance between the morphs and their constituent faces.

If the data obtained matches the results from the current study, it will provide more support for the interaction of morphed faces within face space, as theorised in the current study.

4.9 Conclusions

This study aimed to understand the impact of morphed faces on human face matching performance within a one-to-many face matching task. Overall, this study has shown that untrained humans performed surprisingly well on this task. Compared to a control task, the usage of morphed faces made from 2 constituent faces did not result in any significant difference in performance (accuracy, confidence or response latency) when the target was absent, and actually increased performance in the target present condition. Additionally, using morphed faces made from 8 constituent images improved performance across all levels. However, there was a confound present in the study design, which limits the strength and validity of the findings. Therefore, future research could take steps to prevent such confounds, in addition to comparing the performance against those using morphs made from similar faces, and use multidimensional scaling procedures to locate the morphed and constituent faces within face space. This study provided the foundation for future research to explore the viability of identity fraud through the usage of morphed faces, with the hope of reducing the possibility of such potential incidents in operational settings.

References

Albert, M., Sethuram, A., & Ricanek, K. (2011). Implications of adult facial aging on biometrics. In M. Albert (Ed.), *Biometrics - unique and diverse applications in nature, science, and technology* (pp. 89-106): InTech.

Attorney-General's Department. (n.d.). Face matching services. Retrieved August 28, 2017, from https://www.ag.gov.au/RightsAndProtections/IdentitySecurity/Documents/Fact-Sheet-National-Facial-Biometric-Matching-Capability.pdf

Beveridge, J. R., Zhang, H., Flynn, P. J., Lee, Y., Liong, V. E., Lu, J., . . . Phillips, P. J. (2014, Sept. 29 2014-Oct. 2 2014). *The IJCB 2014 PaSC video face and person recognition competition.* Paper presented at the 2014 IEEE International Joint Conference on Biometrics.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339-360.

Burton, A. M., White, D., & McNeill, A. (2010). The glasgow face matching test. *Behavior Research Methods, 42*(1), 286-291. doi: 10.3758/BRM.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243-248. doi: 10.1111/1467-9280.00144

Busey, T. A. (1998). Physical and psychological representations of faces: Evidence from morphing. *Psychological Science, 9*(6), 476-483.

Busey, T. A., & Tunnicliff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1210-1235. doi: 10.1037/0278-7393.25.5.1210

Butavicius, M. A., Lee, M. D., & Vast, R. L. (2006). Face matching under time pressure and task demands. *Proceedings of the Cognitive Science Society, 28*(28).

Cabeza, R., & Kato, T. (2000). Features are also important: Contributions of featural and configural processing to face recognition. *Psychological Science, 11*(5), 429-433. doi: 10.1111/1467-9280.00283

Calic, D. (2013). *From the laboratory to the real world: Evaluating the impact of impostors, expertise and individual differences on human face matching performance.* (Doctoral thesis), University of Adelaide, Adelaide. Retrieved from https://digital.library.adelaide.edu.au/dspace/handle/2440/91444

Carbon, C.-C. (2011). The first 100 milliseconds of a face: On the microgenesis of early face processing. *Perceptual and Motor Skills, 113*(3), 859-874. doi: 10.2466/07.17.22.PMS.113.6.859-874

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the "uncanny valley hypothesis": Behavioral and functional MRI findings. *Frontiers in Human Neuroscience, 5*, 126. doi: 10.3389/fnhum.2011.00126

Chen, W., Lander, K., & Liu, C. H. (2011). Matching faces with emotional expressions. *Frontiers in Psychology, 2*, 206. doi: 10.3389/fpsyg.2011.00206

Chen, X., Flynn, P. J., & Bowyer, K. W. (2003). Visible-light and infrared face recognition. *ACM Workshop on Multimodal User Authentication*, 48-55.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum Associates.

Department of Foreign Affairs and Trade. (2017). Adult – quick guide to applying for a passport. Retrieved September 25, from https://www.passports.gov.au/passportsexplained/Pages/quicknewadultpassportguide.aspx

Duncan, M. (2006). *A signal detection model of compound decision tasks*. Retrieved from http://cradpdf.drdc-rddc.gc.ca/PDFS/unc65/p528139.pdf.

Ferrara, M., Franco, A., & Maltoni, D. (2014, Sept. 29 2014-Oct. 2 2014). *The magic passport.* Paper presented at the IEEE International Joint Conference on Biometrics, Clearwater, FL.

FISWG. (2012). Guidelines for facial comparison methods. Retrieved August 28, 2017, from https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf

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.), *Avbpa 2003: Audio- and video-based biometric person authentication* (pp. 44-51). Berlin, Heidelberg: Springer Berlin Heidelberg.

Graves, I., Butavicius, M., MacLeod, V., Heyer, R., Parsons, K., Kuester, N., . . . Johnson, R. (2011). The role of the human operator in image-based airport security technologies. In L. C. Jain, E. V. Aidman & C. Abeynayake (Eds.), *Innovations in defence support systems -2: Socio-technical systems* (pp. 147-181). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*(9), 330-337. doi: 10.1016/S1364-6613(00)01519-9

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 986-1004.

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577-596. doi: 10.1080/09658210902976969

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology, 11*(3), 211-222. doi: 10.1002/(SICI)1099-0720(199706)11:3<211::AID-ACP430>3.0.CO;2-O

Kramer, R. S. S., & Ritchie, K. L. (2016). Disguising Superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology, 30*(6), 841-845. doi: 10.1002/acp.3261

Ling, H., Soatto, S., Ramanathan, N., & Jacobs, D. W. (2007, 14-21 October). *A study of face recognition as people age.* Paper presented at the 2007 IEEE 11th International Conference on Computer Vision.

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695-710.

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255-260. doi: 10.1016/S1364-6613(02)01903-4

McCaffery, J. M., & Burton, A. M. (2016). Passport checks: Interactions between matching faces and biographical details. *Applied Cognitive Psychology, 30*(6), 925-933. doi: 10.1002/acp.3281

McIntyre, A. H., Hancock, P. J. B., Kittler, J., & Langton, S. R. H. (2013). Improving discrimination and face matching with caricature. *Applied Cognitive Psychology, 27*(6), 725-734. doi: 10.1002/acp.2966

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*(4), 865-876. doi: 10.3758/BF03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69*(7), 1175-1184. doi: 10.3758/BF03193954

Megreya, A. M., & Burton, M. A. (2008). Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14*(4), 364-372. doi: 10.1037/a0013464

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

Mi, J.-X., Li, C., Li, C., Liu, T., & Liu, Y. (2016). A human visual experience-inspired similarity metric for face recognition under occlusion. *Cognitive Computation, 8*(5), 818-827. doi: 10.1007/s12559-016-9420-x

Nestor, A., Plaut, D. C., & Behrmann, M. (2013). Face-space architectures. *Psychological Science, 24*(7), 1294-1300. doi: 10.1177/0956797612464889

Phillips, P. J., Beveridge, J. R., Draper, B. A., Givens, G., O'Toole, A. J., Bolme, D. S., . . . Weimer, S. (2011). *An introduction to the good, the bad, and the ugly face recognition challenge problem.* Paper presented at the Ninth IEEE International Conference on Automatic Face and Gesture Recognition, 2011.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(10), 1090-1104. doi: 10.1109/34.879790

Phillips, P. J., & O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing, 32*(1), 74-85. doi: 10.1016/j.imavis.2013.12.002

Phillips, P. J., 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. doi: 10.1016/S0262-8856(97)00070-X

Ricanek, K., & Tesafaye, T. (2006, 2-6 April 2006). *Morph: A longitudinal image database of normal adult age-progression.* Paper presented at the IEEE 7th International Conference of Automatic Face and Gesture Recognition and Workshops, Southampton, UK.

Robertson, D. J., Kramer, R. S. S., & Burton, A. M. (2017). Fraudulent ID using face morphs: Experiments on human and automatic recognition. *PLoS ONE, 12*(3), e0173319. doi: 10.1371/journal.pone.0173319

Schretlen, D. J., Pearlson, G. D., Anthony, J. C., & Yates, K. O. (2001). Determinants of Benton facial recognition test performance in normal adults. *Neuropsychology, 15*(3), 405-410.

Schwaninger, A., Lobmaier, J. S., & Collishaw, S. M. (2002). Role of featural and configural information in familiar and unfamiliar face recognition. In H. H. Bülthoff, C. Wallraven, S.-W. Lee & T. A. Poggio (Eds.), *Biologically motivated computer vision: Second international workshop, BMCV 2002* (Vol. 2525, pp. 643-650). Berlin, Heidelberg: Springer Berlin Heidelberg.

Seyama, J. i., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators & Virtual Environments, 16*(4), 337-351.

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence-accuracy relationship in unfamiliar face matching. [Advance online publication]. *Journal of Experimental Psychology: Applied*. doi: 10.1037/xap0000130

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*(1), 47-58. doi: 10.1037/xap0000108

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*(2), 161-204. doi: 10.1080/14640749108400966

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *The Quarterly Journal of Experimental Psychology, 69*(10), 1996-2019. doi: 10.1080/17470218.2014.990392

Van der Linde, I., & Watson, T. (2010). A combinatorial study of pose effects in unfamiliar face recognition. *Vision Research, 50*(5), 522-533. doi: 10.1016/j.visres.2009.12.012

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS ONE, 10*(10), e0139827. doi: 10.1371/journal.pone.0139827

White, D., Kemp, R. I., Jenkins, R., & Burton, A. M. (2014). Feedback training for facial image comparison. *Psychonomic Bulletin & Review, 21*(1), 100-106. doi: 10.3758/s13423-013-0475-3

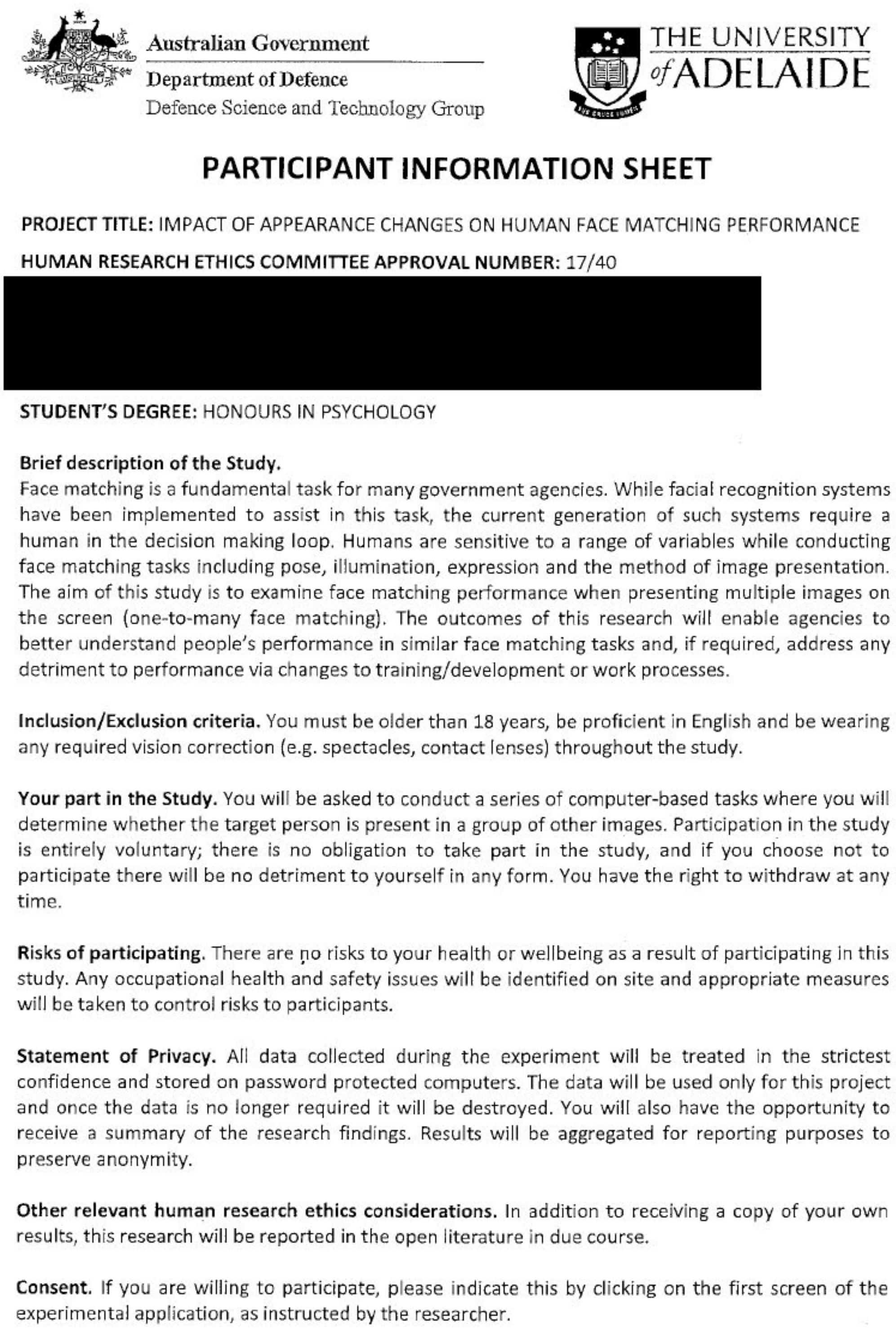
White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PLoS ONE, 9*(8), e103510. doi: 10.1371/journal.pone.0103510

White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences, 282*(1814), 20151292. doi: 10.1098/rspb.2015.1292

Wirth, B. E., & Carbon, C.-C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied, 23*(2), 138-157. doi: 10.1037/xap0000114

Appendices

Appendix A: Participant Information Sheet



Australian Government

Department of Defence

Defence Science and Technology Group

THE UNIVERSITY *of* ADELAIDE

PARTICIPANT INFORMATION SHEET

PROJECT TITLE: IMPACT OF APPEARANCE CHANGES ON HUMAN FACE MATCHING PERFORMANCE

HUMAN RESEARCH ETHICS COMMITTEE APPROVAL NUMBER: 17/40

[redacted]

STUDENT'S DEGREE: HONOURS IN PSYCHOLOGY

Brief description of the Study.

Face matching is a fundamental task for many government agencies. While facial recognition systems have been implemented to assist in this task, the current generation of such systems require a human in the decision making loop. Humans are sensitive to a range of variables while conducting face matching tasks including pose, illumination, expression and the method of image presentation. The aim of this study is to examine face matching performance when presenting multiple images on the screen (one-to-many face matching). The outcomes of this research will enable agencies to better understand people's performance in similar face matching tasks and, if required, address any detriment to performance via changes to training/development or work processes.

Inclusion/Exclusion criteria. You must be older than 18 years, be proficient in English and be wearing any required vision correction (e.g. spectacles, contact lenses) throughout the study.

Your part in the Study. You will be asked to conduct a series of computer-based tasks where you will determine whether the target person is present in a group of other images. Participation in the study is entirely voluntary; there is no obligation to take part in the study, and if you choose not to participate there will be no detriment to yourself in any form. You have the right to withdraw at any time.

Risks of participating. There are no risks to your health or wellbeing as a result of participating in this study. Any occupational health and safety issues will be identified on site and appropriate measures will be taken to control risks to participants.

Statement of Privacy. All data collected during the experiment will be treated in the strictest confidence and stored on password protected computers. The data will be used only for this project and once the data is no longer required it will be destroyed. You will also have the opportunity to receive a summary of the research findings. Results will be aggregated for reporting purposes to preserve anonymity.

Other relevant human research ethics considerations. In addition to receiving a copy of your own results, this research will be reported in the open literature in due course.

Consent. If you are willing to participate, please indicate this by clicking on the first screen of the experimental application, as instructed by the researcher.

Contact details.
Should you have any complaints or concerns about the manner in which this project is conducted, please do not hesitate to contact the investigators in person.

[redacted]

Alternatively, you may contact Professor Paul Delfabbro, the convenor of the School of Psychology Human Ethics Sub-Committee.

School of Psychology Human Ethics Sub-Committee Convenor
Professor Paul Delfabbro
The University of Adelaide School of Psychology
Hughes Building (Room 410)
ADELAIDE SA 5005
Telephone: (08) 8313 3770
Email: paul.delfabbro@adelaide.edu.au

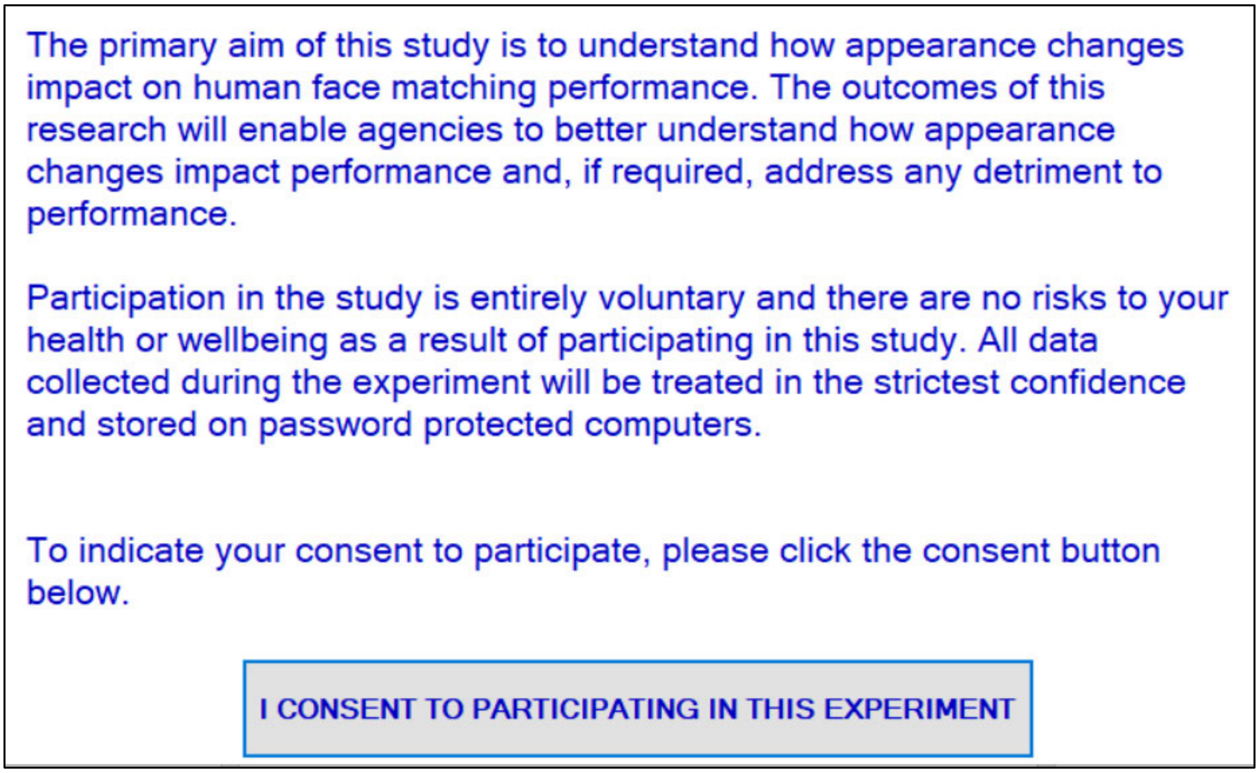
Contact can also be made with the DST Group Ethics Review Panel.

Chair, DST Group Ethics Review Panel
Dr Ken McAnally
Aerospace Division
506 Lorimer St
Fishermans Bend VIC 3207
Telephone: (03) 9626 7251
Email: HumanSciencesEthics@dsto.defence.gov.au

Issues remaining following discussion with the DST Group Ethics Review Panel may be discussed with the Executive Secretary of the Australian Defence Human Research Ethics Committee.

Executive Secretary
Australian Defence Human Research Ethics Committee
CP2–7–124
Department of Defence
CANBERRA ACT 2600
Telephone: (02) 6266 3837
Facsimile: (02) 6266 4982
Email: ADHREC@defence.gov.au

Appendix B: Consent



The primary aim of this study is to understand how appearance changes impact on human face matching performance. The outcomes of this research will enable agencies to better understand how appearance changes impact performance and, if required, address any detriment to performance.

Participation in the study is entirely voluntary and there are no risks to your health or wellbeing as a result of participating in this study. All data collected during the experiment will be treated in the strictest confidence and stored on password protected computers.

To indicate your consent to participate, please click the consent button below.

Appendix C: Study Advertisement

DOES THE FACE ON THE LEFT BELONG TO ANYONE ON THE RIGHT?

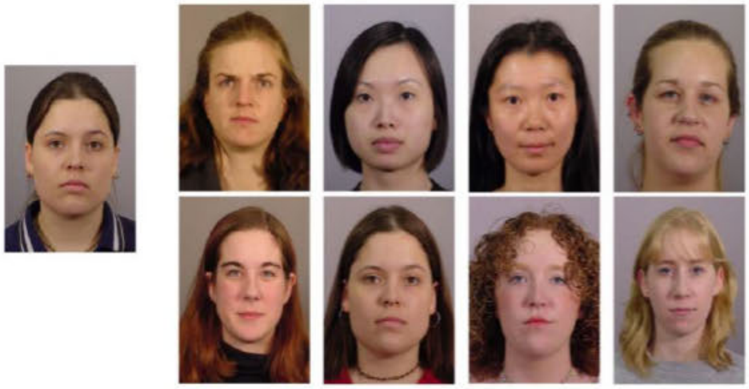


Image from Vu, L. (2016). ***The Impact of Plastic Surgery on One-to-Many Human Face Matching Performance***
(Honours thesis, University of Adelaide, Australia)

Did you find this interesting? Do you want to further test your skills?

We're searching for potential participants to search another 112 pairs of faces, to help us understand how people perform in similar tasks.

<u>What will you get out of this?</u>

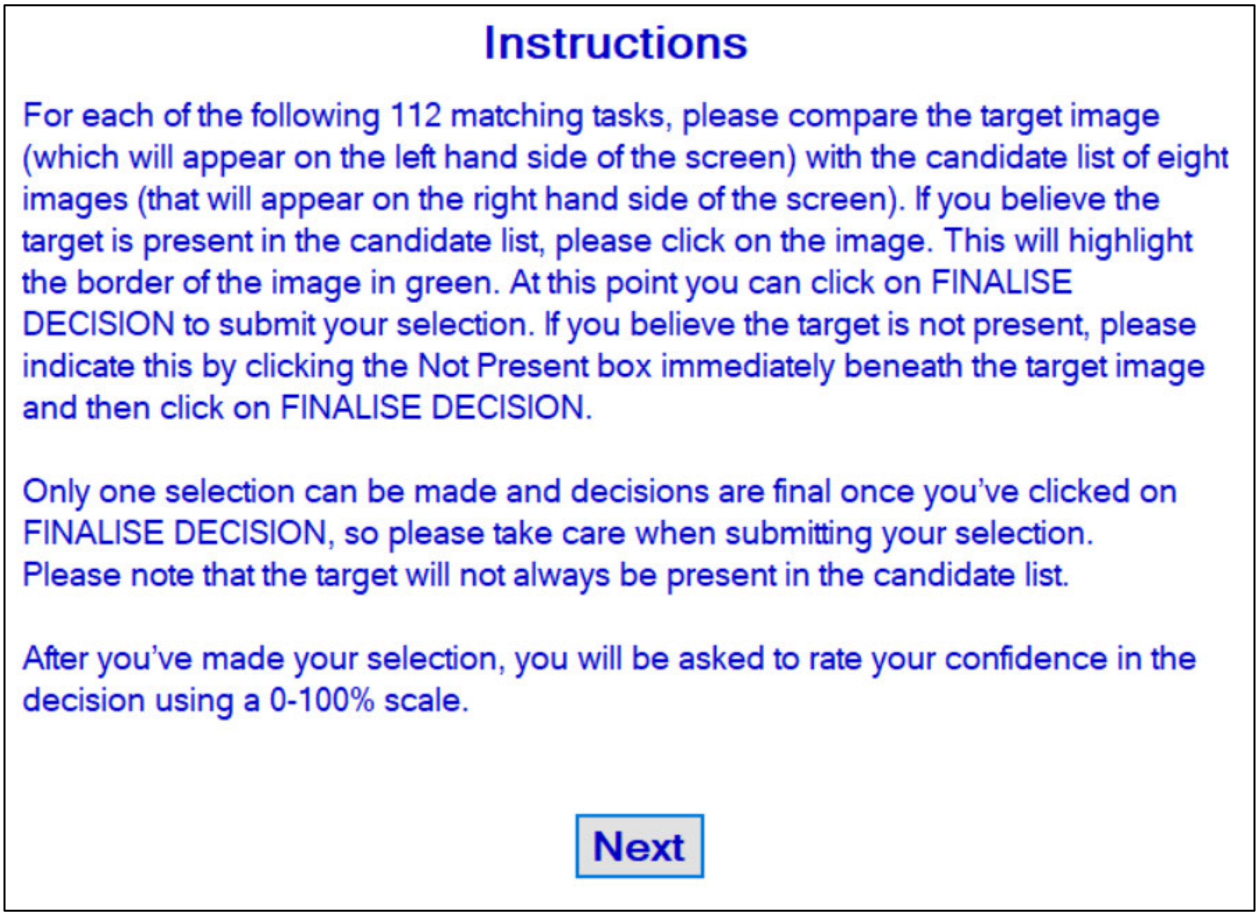
- Refreshments
- A copy of your results (if you want it), so you can see how well you performed
- Knowledge that you have contributed to research, that will help the development of better facial recognition systems and training for a range of government agencies

<u>Inclusion criteria</u>

- Older than 18 years
- Proficient in English
- Bring along any required vision correction (e.g. spectacles, contact lenses) to the study

[redacted]

Appendix D: Instruction Screen Before Practice Trial



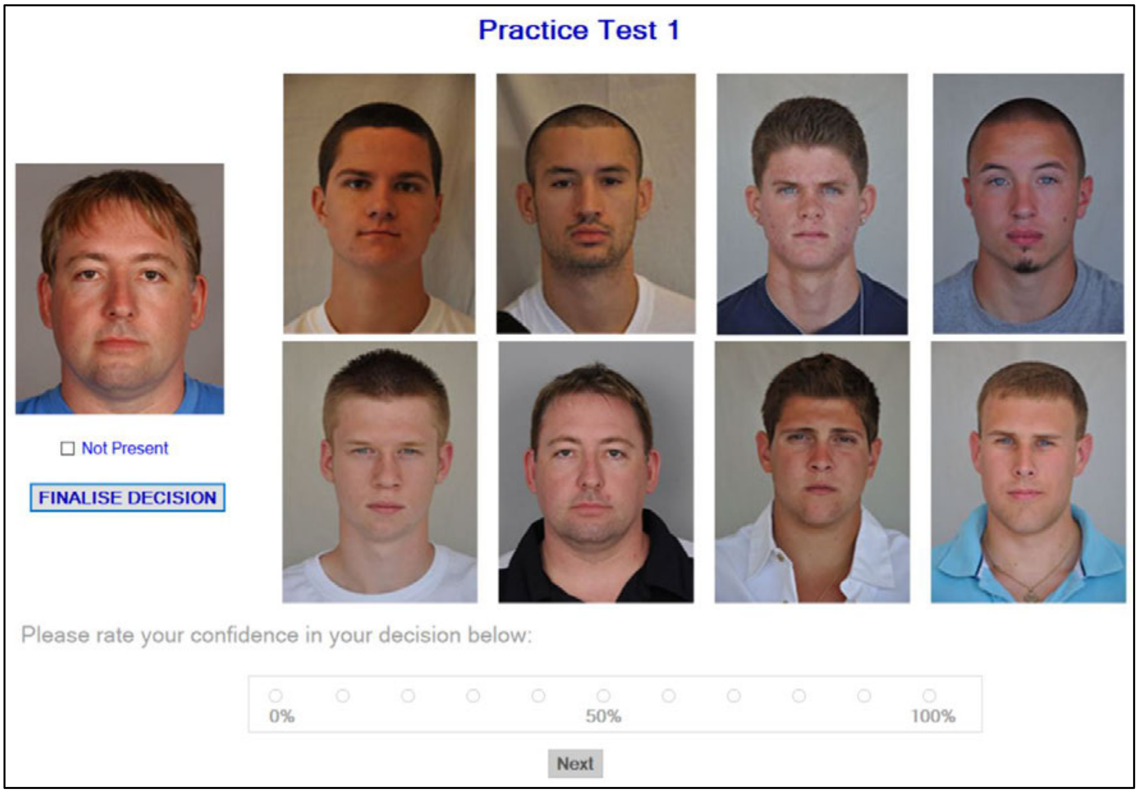
Instructions

For each of the following 112 matching tasks, please compare the target image (which will appear on the left hand side of the screen) with the candidate list of eight images (that will appear on the right hand side of the screen). If you believe the target is present in the candidate list, please click on the image. This will highlight the border of the image in green. At this point you can click on FINALISE DECISION to submit your selection. If you believe the target is not present, please indicate this by clicking the Not Present box immediately beneath the target image and then click on FINALISE DECISION.

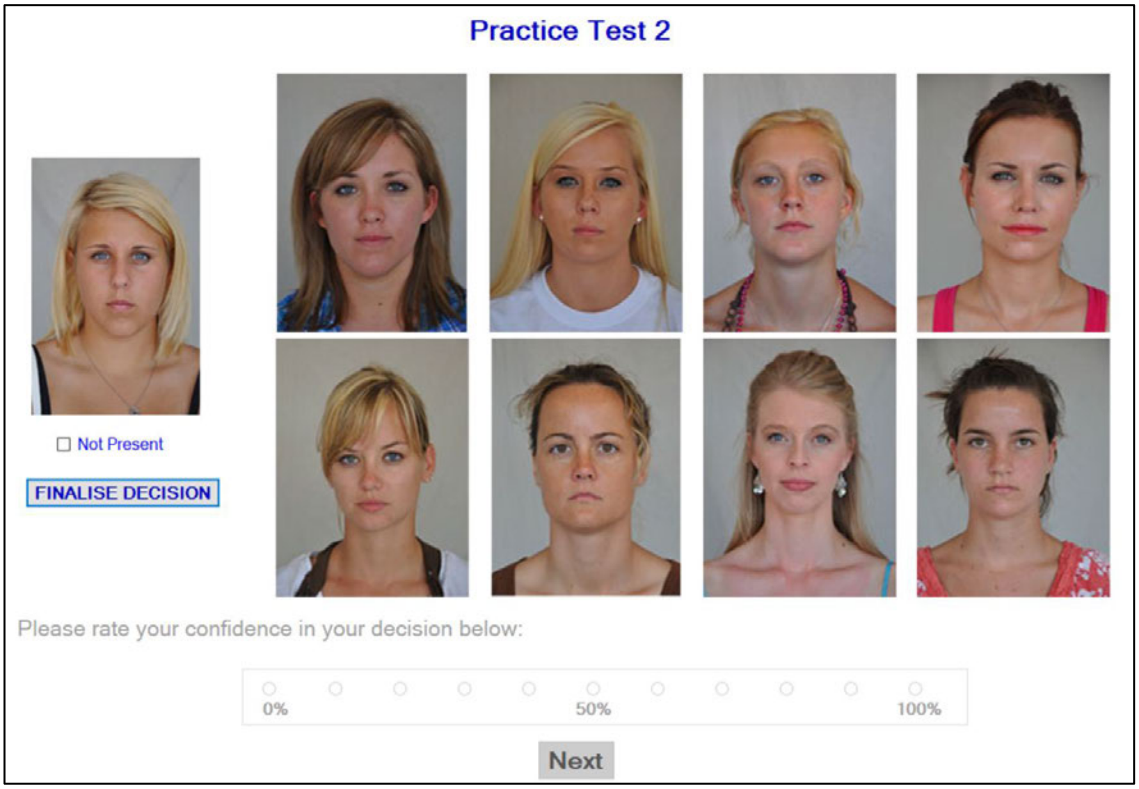
Only one selection can be made and decisions are final once you've clicked on FINALISE DECISION, so please take care when submitting your selection.
Please note that the target will not always be present in the candidate list.

After you've made your selection, you will be asked to rate your confidence in the decision using a 0-100% scale.

Appendix E: Practice Trials


Practice Test 1

Please rate your confidence in your decision below:


Practice Test 2

Please rate your confidence in your decision below:

Appendix F: Testing Normality with Shapiro-Wilk Tests

Variable	*W*	*P*	Skewness	Kurtosis
Accuracy				
2 Images Overall	.90	< .001	-0.88	-0.35
8 Images Overall	.71	< .001	-2.39	7.74
16 Images Overall	.81	< .001	-2.00	6.79
Control Overall	.88	< .001	-1.06	0.42
Hit				
2 Images	.59	< .001	-2.74	9.17
8 Images	.51	< .001	-3.21	13.02
16 Images	.84	< .001	-0.57	-0.55
Control	.84	< .001	-1.11	1.37
False alarm				
2 Images	.86	< .001	1.02	0.09
8 Images	.63	< .001	2.93	11.18
16 Images	.61	< .001	3.08	12.19
Control	.84	< .001	1.15	0.56
Confidence				
2 Images Overall	.98	.440	-0.50	0.66
8 Images Overall	.95	.040	-0.75	0.82
16 Images Overall	.96	.088	-0.38	0.04
Control Overall	.94	.016	-0.76	1.63
2 Images Target Present	.92	.002	-0.97	0.73
8 Images Target Present	.89	< .001	-1.00	0.33
16 Images Target Present	.96	.065	-0.65	0.32
Control Target Present	.94	.010	-0.94	1.52
2 Images Target Absent	.88	< .001	-1.73	5.84
8 Images Target Absent	.97	.134	-0.59	0.75
16 Images Target Absent	.98	.433	-0.18	-0.42
Control Target Absent	.96	.111	-0.57	0.99

Variable	*W*	*p*	Skewness	Kurtosis
Response Latency				
2 Images Overall	.89	< .001	1.43	2.58
8 Images Overall	.97	.129	0.42	-0.49
16 Images Overall	.95	.040	0.72	0.32
Control Overall	.96	.099	0.52	-0.25
2 Images Target Present	.83	< .001	2.04	5.99
8 Images Target Present	.94	.008	0.82	0.15
16 Images Target Present	.91	.001	1.15	1.37
Control Target Present	.94	.010	0.96	1.14
2 Images Target Absent	.87	< .001	1.77	5.07
8 Images Target Absent	.97	.312	0.46	-0.30
16 Images Target Absent	.93	.004	0.94	0.84
Control Target Absent	.95	.027	0.75	0.22

Note. *df* = 52 for all analyses. *SE* = 0.33 for all skewness output, while *SE* = 0.65 for all kurtosis output