Received: 12 September 2022 | Revised: 20 June 2023 | Accepted: 26 June 2023

DOI: 10.1002/acp.4110

RESEARCH ARTICLE

WILEY

Improving face morph detection with the pairs training effect

Tessa R. Flack | Kay L. Ritchie | Charlotte Cartledge | Elizabeth A. Fuller | Robin S. S. Kramer

School of Psychology, University of Lincoln, Lincoln, UK

Correspondence
Tessa R. Flack, School of Psychology, University of Lincoln, Lincoln, LN6 7TS, UK.
Email: tflack@lincoln.ac.uk

Abstract

It is becoming increasingly common for face morphs (weighted combinations of two people's photographs) to be submitted for inclusion in an official document, such as a passport. These images may sufficiently resemble both individuals that they can be used by either person in a 'fraudulently obtained genuine' document. Problematically, people are poor at detecting face morphs and there is limited evidence that this can be improved. Here, we tested whether the 'pairs training effect' (working in pairs, which we know improves unfamiliar face matching) can improve face morph detection. We found morph detection was more accurate when working in a pair. Further, the lower performer in the pair maintained this benefit when completing the task again individually. We conclude that the pairs training effect translates to face morph detection, and these findings have important implications for improving the detection of face morphs at the initial application stage.

KEYWORDS
face morph, face perception, morphing attack, pairs

1 | INTRODUCTION

Face photographs make up an important part of many forms of official documentation, such as passports and driving licences. We use these images to make important decisions, such as whether someone should be allowed to enter a country. An increasingly common type of ID fraud is a 'face morph attack' (Ferrara et al., 2014). These utilise specialist software to combine images of two different people into a single morph. More complex than a simple average of the two images, this process typically involves applying landmarks to both facial images and then generating a set of intermediate frames as one image is transformed into the other (for more detail, see Ferrara & Franco, 2022). The choice of frame to use will depend on the preferred weighting of the two original images in the final morph, with the goal that this morph sufficiently resembles both individuals that it can be used by either person. For example, Person A (who has no criminal record) creates a morphed photo combining themselves and Person B (whose criminal record prevents them from travelling internationally). Person A submits the morphed image as their new photo when renewing their passport and the morphed image is compared with the previous images on record of Person A. As the morph sufficiently resembles Person A, the image is accepted, and a new passport is issued using the morph. Person B is then able to successfully use the passport to travel since the image also sufficiently resembles them. This type of fraud is difficult to detect as the resulting document is genuine. With the advancement in image manipulation software, face morphs can be of such high quality that they are especially difficult to detect (Ferrara et al., 2016; Kramer et al., 2019).

1.1 | Key issues in face morph detection

As previously mentioned, given that the resulting document is genuine, standard anti-counterfeit measures (e.g., the use of security

watermarks) are all present with this type of fraud. Detection, therefore, must take place at the point of issuance (comparing the morph image to the previously stored face image) or presentation (comparing the morph image to the individual's 'live' face at border control). There are two main directions for improving detection of face morph images: (1) use computer algorithms to detect face morph images; and (2) train human operators to accurately detect face morph images.

One key problem with using face images in identification documents is that matching an unfamiliar individual's face to a photograph is surprisingly difficult. When an official, for example a border control officer, is presented with a photo-ID, they are typically required to decide whether the face image in the photo-ID matches the face of its carrier. It has been widely demonstrated that people are poor at matching unfamiliar faces (Bruce et al., 2001; Burton et al., 2010; Jenkins et al., 2011; Ritchie et al., 2015), and performance does not appear to improve with experience (White, Burton, et al., 2014). Therefore, when it comes to matching an individual's face to another image of that individual, this process is surprisingly error prone (White, Burton, et al., 2014). A wide range of research has demonstrated that those with professions that frequently require this type of unfamiliar face matching task, such as police officers, passport officers, and supermarket cashiers, are no better than undergraduate students (Burton et al., 1999; Kemp et al., 1997; White, Kemp, et al., 2014). Therefore, detecting a face morph at the point of presentation is likely to be more difficult when combined with the task of matching the face to the individual and when the morph image is already incorporated into an official document. Most importantly, production of fraudulent documents should be avoided in the first instance. Therefore, resources should be focused on detecting face morph images at the point of issuance.

1.2 | The role of computer algorithms in face morph detection

As the quality of face morphs increases, they are likely to become even more difficult to detect. A growing body of literature has shown there is potential to develop increasingly sophisticated computer algorithms that can detect face morphs (Makrushin et al., 2017; Neubert, 2017; Raghavendra et al., 2017a, 2017b; Scherhag et al., 2017, 2019, 2022; Seibold et al., 2018; Venkatesh et al., 2021). For example, algorithms are able to detect the inconsistencies left behind by the morphing process, which are not easily detected by a human observer, such as inconsistences in the reflections across the image (Seibold et al., 2018). The introduction of such algorithms would require a systematic shift in process for the passport office and/or border control. In addition, although a significant volume of research is being dedicated to creating computer algorithms that can reliably detect face morphs, there are some key challenges hindering the development of an algorithm that can produce robust and reliable performance in a real-life border control or passport issuing office scenario (Ferrara & Franco, 2022; Scherhag et al., 2022; Venkatesh et al., 2021). For example, a lack of availability of large datasets for training, and databases including both digital and print images, means that algorithm developers can only train on what is available to them. Often this is a limited set of images that may have been created in the same way. This leads to an overoptimistic detection rate (Scherhag et al., 2022). In addition, the large databases that are publicly accessible are only available to test morph attack detection systems rather than to be used for training the algorithms (Venkatesh et al., 2021).

Generalisability is also a key issue for morph detection algorithms. Although many morph attack detection methods perform very well with specific, constrained image databases, these systems are vulnerable when applied to different sets of images (e.g., that were created in a different way) or when used to detect the types of images seen in real-world morph attack settings (Spreeuwers et al., 2022). For example, an algorithm in a practical border control scenario needs to be robust to print-scan transformations, where the images have been scanned, resized, compressed, and printed, as part of the passport application process (Spreeuwers et al., 2022; Venkatesh et al., 2021). Therefore, widescale implementation of a morph detection algorithm is unlikely in the near future, and so a more easily implemented and cost-effective solution is required in the interim.

1.3 | Can training improve human face morph detection?

Previous research has suggested it is possible to train individuals to detect face morphs (Robertson et al., 2017, 2018). Robertson et al. (2017) presented participants with pairs of images. The participants were asked to take on the role of a passport officer and decide whether a traveller matched the photograph in their passport. When participants were informed about face morphs and their use in creating fraudulent ID documents, participants performed well, with 50/50 morphs (weighting both identities equally) accepted as a match only 21% of the time (compared to 68% when participants were not informed about morphs). In a later study, Robertson et al. (2018) used a morph detection training task, where participants were presented with pairs of faces—one morph and one original image. Participants indicated which of the two images was a morph, with feedback provided on each trial, followed by a short period where the participants could examine the morph. Participants completed a morph detection task pre- and post-training. Participants who received morph detection training saw a significant improvement from their baseline detection performance (which was at chance level) with overall detection rates rising to nearly 80%. Further analysis revealed that training was only effective for the lowest performing participants in the baseline morph detection task, and training did not improve morph detection rates for participants who performed well initially.

Although the research described above showed a promising role for training in morph detection, at least for those who initially performed poorly, further investigation using higher quality morphs puts this conclusion into question. The quality of face morph images is rapidly increasing with the wider availability of sophisticated image editing software. The standard technique for creating morphs uses

landmark-based morph generation, where points are placed on regions of the face, such as the eyes, nose, mouth, and jaw line. The points from both faces are warped by moving the pixels to different, more averaged positions (Venkatesh et al., 2021). There are a variety of techniques used, but all processes translate both the landmarks and the associated texture of the faces. When using this technique, there are often artefacts in the resulting image caused by misaligned pixels, some of which can be very noticeable to a human observer, for example, double-edge effects, where edges are erroneously repeated. Therefore, after morph creation, manual touch-ups are often required to remove any visible artefacts. Problematically, the images used in Robertson et al. (2017, 2018) still contained visible artefacts from the morphing process (e.g., a ghostly second hair line).

Later studies have used higher quality morphs, more likely to be representative of the morphs created currently by fraudsters (Kramer et al., 2019; Nightingale et al., 2021). Kramer et al. (2019) replicated Robertson et al.'s (2018) study with higher quality morphs, where any artefacts of the morphing process were removed in a manual post-processing stage. In this study, participants who received morph detection training did not improve from baseline performance (which was very poor). In a further experiment, Kramer et al. (2019) used a forced choice paradigm, where participants were presented with one image per trial and simply asked to indicate whether they believed the image was a morph or not. Participants were either presented with morph detection guidance (akin to that used by Robertson and colleagues) followed by the morph detection task, or only completed the morph detection task (control group). Performance was again poor, with 54% and 57% accuracy for the control and morph detection guidance groups respectively. Nightingale et al. (2021) conducted a similar study with a larger and more diverse stimuli set. Mean accuracy at the morph detection task was 54%, although after training similar to that of Kramer et al. (2019), accuracy rose to 60%. Clearly, when higher quality morphs are used, morph detection is a very difficult task and the scope for improvement via training appears limited.

1.4 | The pairs training effect

One method of training that has been used to improve another type of face task (unfamiliar face matching), is to have participants complete the task in pairs. The 'pairs training effect' has been shown to produce consistent and reliable improvements in unfamiliar face matching, where participants are asked to decide whether two images are of the same person or two different people (Dowsett & Burton, 2015; Ritchie et al., 2022). The design sees participants completing three blocks of an unfamiliar face matching task, where all blocks are of equal difficulty. At Time 1, the participants complete the face matching task individually. At Time 2, they complete the task as a pair and are explicitly told to discuss the stimuli and come to a joint decision for each trial. At Time 3, the participants complete the remaining block of the face matching task individually. Participants perform significantly better in pairs, and those who were the lower performers (in each pair) at Time 1 show better performance at Time 3 compared to their baseline Time 1 performance, suggesting they have learned something from working in the pair. A recent study demonstrated this pairs training effect to be replicable, and the training effect itself is even maintained after a delay (Ritchie et al., 2022). This effect was not driven by practice effects, but by the interaction that occurs within the pairs.

1.5 | The current studies

As noted above, the implementation of computer algorithms able to detect face morphs is likely to be some way off and would require a large, systematic change in procedures. This highlights the need for a simple and easy-to-implement solution that can be used in the interim, and provides the opportunity to apply training, already seen to improve performance in other types of face perception tasks, to face morph detection. Here, we investigated whether the pairs training effect, which shows consistent and reliable improvements in unfamiliar face matching, could also improve face morph detection. In the initial task creation experiment (Experiment 1), we created a face morph detection task comprising three blocks of equal difficulty. We provided participants with face morph detection guidance in line with previous studies (Kramer et al., 2019; Nightingale et al., 2021; Robertson et al., 2017, 2018) and tested whether performance increased as a result of simple practice at the task. In the main experiment, we used our new face morph detection task in a pairs training effect paradigm to assess whether working in a pair could improve face morph detection.

2 | EXPERIMENT 1: TASK CREATION AND CONTROL EXPERIMENT

Our first experiment established the stimuli and task for the main pairs training experiment. It is essential for the pairs training effect that we use three blocks of equal difficulty so that any observed improvements will unlikely be due to differences in item difficulty across blocks (see Dowsett & Burton, 2015; Ritchie et al., 2022). This experiment also acted as a control, ruling out practice effects as an explanation for any pairs effects observed in the main experiment. Finally, we sought to ensure that simply pairing the participants artificially during analysis did not lead to any statistical artefacts that could be misinterpreted as a pairs training effect. Participants completed three blocks of a face morph detection task individually, responding on each trial as to whether the image was a morph or not.

2.1 | Method

2.1.1 | Participants

Seventy-two participants were recruited via Prolific (2022) (www.prolific.co) to take part (42 female, 62 self-reported White, mean age 30.6 years ± 10.3 SD, range 18–56). Our sample size was informed by Ritchie et al.'s (2022) control experiment, which used 50 participants.

All participants gave informed consent and ethical approval was granted by the School of Psychology Research Ethics Committee at the University of Lincoln.

2.1.2 | Stimuli

Stimuli were taken from a previous study (Kramer et al., 2019) and depicted front-facing student models with neutral expressions. Posture, lighting, and distance to the camera were kept constant, and glasses and jewellery were removed. To create the morphs, models were paired with each other based on general descriptors (e.g., male with brown hair). JPsychomorph (Tiddeman et al., 2001) was used to create the morphs, and Adobe Photoshop was used to remove any noticeable artefacts from the averaging process (see Figure 1). Finally, images were resized to 440 × 570 pixels. For additional details on morph creation, see Kramer et al. (2019). These morph images were provided to the National Institute of Standards and Technology, US Dept of Commerce, and were included in the Face Recognition Vendor Test MORPH. This test evaluates morph detection performance for face morph attack detection algorithms. There was a high morph miss rate for our morphs for many of the algorithms tested (National Institute of Standards and Technology, 2020). Therefore, these easily created morphs can fool state-of-the-art algorithms.

Ninety images were selected for the current study: 45 morphs and 45 exemplars. These exemplars (unaltered, original images) depicted identities not used in the creation of the morphs. Based on accuracy data from Kramer et al. (2019), we determined trial difficulty for all trials (morphs and exemplar images) by calculating the average accuracy across participants for each trial. We then constructed three blocks of equal difficulty (Block 1: $M = 52.12\%$, Block 2: $M = 52.97\%$, and Block 3: $M = 51.81\%$) with 30 trials in each (15 morphs, 15 exemplars).

2.1.3 | Procedure

The study was administered using the online platform Qualtrics (2022) (Qualtrics, Provo, UT). Before beginning the morph detection task, as in Kramer et al. (2019), participants were presented with 'Morph Fraud/Detection Tips'. These onscreen tips explained that a morph is produced when images of two different people have been averaged together by computer software, and showed two exemplars and the resulting morph created from the two images. In addition, the tips screen gave advice on potential ways to identify morphs: (1) 'morphs often have smother skin than normal photographs'; and (2) 'morphs may also show irregularities within the hair's texture'. The experiment comprised three blocks of a morph detection task, with 30 trials per block (15 morphs, 15 exemplars). In line with Dowsett and Burton (2015) blocks were presented in a fixed order, with images within each block presented in a random order. For each trial, participants were asked, 'Is this image a face morph?' Participants responded by selecting either 'Yes' or 'No' from the onscreen options. No feedback was given.

2.2 | Results and discussion

This study fulfilled two purposes: (1) to validate the task and ensure blocks were of equal difficulty, and (2) to rule out that any pairs training effect observed in the main experiment was caused by practice effects or any statistical artefacts resulting from simply pairing participants and analysing accordingly. To create the artificial pairings, participants were grouped into pseudo-pairs by considering participants 1 and 2 as a pair, participants 3 and 4 as a pair, and so on. As in previous studies (Dowsett & Burton, 2015; Ritchie et al., 2022), pairs were split into high and low performers based on their performance at Time 1. For four of the pairs, both participants had equal accuracy (percentage correct) at Time 1 and so were separated into high and low performers based on sensitivity (d').

Accuracy data were entered into a 2 × 3 mixed ANOVA (Kaufmann & Schering, 2007) with the factors Pair Member (high, low; between-subjects) and Block (T1, T2, T3; within-subjects). There was a significant main effect of Pair Member, with higher accuracy seen for the high performers, $F(1, 70) = 17.03$, $p < .001$, $\eta_p^2 = .20$, $BF_{10} > 100$. There was also a significant main effect of Block, $F(2, 140) = 3.63$, $p = .029$, $\eta_p^2 = .05$, $BF_{10} = 0.95$, however no pairwise comparisons survived Bonferroni correction (T1 vs. T2, $p = .052$, T1 vs. T3, $p = .065$, T2 vs. T3, $p = 1.00$). This suggests there was no reliable difference in the difficulty of the three blocks, meaning any later effects seen in Experiment 2 are unlikely to be driven by

FIGURE 1 An example of the images used in the current experiments. The faces on the left and right depict two different individuals, and the face in the centre is a 50/50 morph of these two individuals. The individuals pictured have given permission for their images to be reproduced here. Image adapted from Kramer et al. (2019), fig. 1.

differences in item difficulty across the blocks. Performance on the task (Block 1, $M = 60.2\%$; Block 2, $M = 56.8\%$; Block 3, $M = 57.2\%$) was equivalent to that reported in Kramer et al.'s (2019) Experiment 2, where participants completed the same task ($M = 57.1\%$). In addition, performance did not increase across the three blocks, indicating there was no practice effect from completing multiple blocks of the task.

There was no significant Pair Member × Block interaction, $F(2, 140) = 1.46$, $p = .237$, $\eta_p^2 = .02$, $BF_{10} = 0.59$, suggesting that a pairs training effect does not occur simply due to pseudo-pairing of the data (see Figure 2). This lack of an interaction demonstrates that, although the high and low performers are (necessarily) different at T1, there is no change in this difference across the experimental sessions, which we might expect if regression to the mean played a role. As such, we can be confident that any effects found in Experiment 2 are unlikely the result of this statistical explanation.

3 | EXPERIMENT 2: PAIRS EXPERIMENT

Our second experiment addressed whether the pairs training effect paradigm, previously successful in unfamiliar face matching, could be used to improve performance on another type of face task—face morph detection. Participants completed the morph detection task created in Experiment 1, but this time they were recruited and participated in pairs. Participants completed Time 1 alone, but then completed Time 2 with their partner. As in the original studies featuring the pairs training effect paradigm, participants were told to discuss and come to a joint decision on each trial. Participants then completed Time 3 alone again. Based on previous research on unfamiliar face matching (Dowsett & Burton, 2015; Ritchie et al., 2022), we expected to see an increase in face morph detection at Time 2, when participants worked together. In addition, we predicted that the lower performer in the pair would not only perform better while in the pair, but would also perform better post-pair (Time 3) compared to their initial baseline performance (Time 1).

3.1 | Method

3.1.1 | Participants

Ninety-four participants (47 pairs) took part in Experiment 2 (62 female, 92 self-reported White, mean age 30.3 years ± 14.6 SD, range 18–70). Our sample size was informed by Ritchie et al. (2022). All participants gave informed consent and ethical approval was granted by the School of Psychology Research Ethics Committee at the University of Lincoln.

3.1.2 | Stimuli and procedure

Stimuli and procedure were the same as in Experiment 1, with the following exceptions. Participants were recruited and participated in pairs, with pairs comprising two friends or relatives who were familiar with one another. Due to the COVID-19 pandemic, participants were tested online and not under researcher supervision. However, recruitment was targeted to ensure reliable participants, and instructions and guidance were made clear and trialled to ensure comprehension. Participants completed Time 1 alone, with the other participant instructed to sit away from the device and not to communicate or look at the other participant's screen. After both participants had completed Time 1, participants sat together at the same device to complete Time 2, and were instructed to discuss each trial and come to a joint decision. Participants then completed Time 3 individually (again, sitting away from each other and refraining from any communication). Accuracy data from Time 1 and Time 3 suggest that participants did indeed follow instructions and complete these blocks separately, as performance between pair members was distinguishable. Had participants not

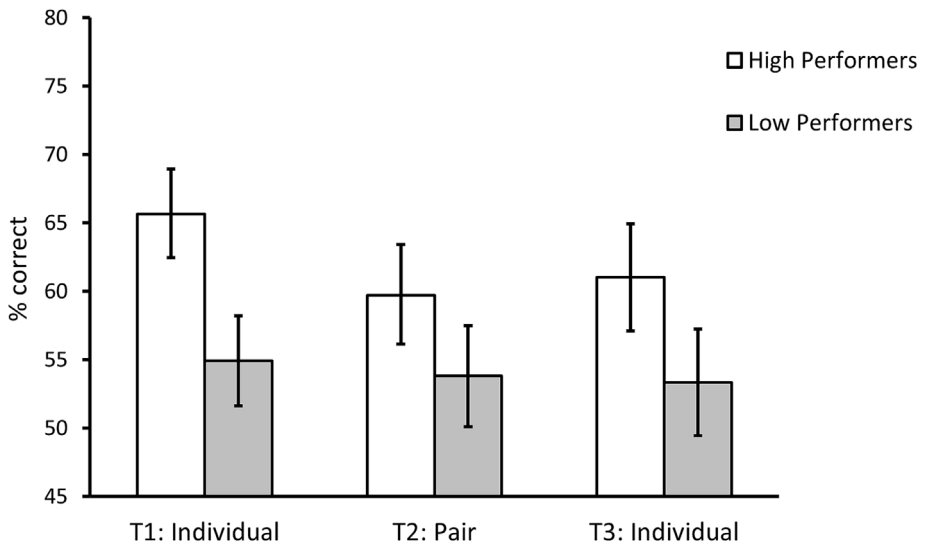

FIGURE 2 Accuracy data for the morph detection task in Experiment 1, artificially pairing participants based on their T1 performance. Error bars represent 95% confidence intervals of the mean.

followed the instructions and completed these together, this would serve to weaken any observed effect.

3.2 | Results

3.2.1 | Assessing the pairs training effect

In order to determine whether working in a pair improved morph detection, accuracy data were entered into a one-way repeated measures ANOVA with the factor Experimental Session (T1 [individual], T2 [pair], and T3 [individual]). There was a significant main effect of Experimental Session, $F(2, 186) = 6.86$, $p = .001$, $\eta_p^2 = .07$, $BF_{10} = 17.28$. Bonferroni corrected pairwise comparisons showed significantly higher accuracy at T2 ($M = 61.91\%$) compared to T1 ($M = 57.70\%$, $p = .008$) and T3 ($M = 58.09\%$, $p = .008$). Accuracy at T1 and T3 were not significantly different ($p = 1.00$).

In line with Dowsett and Burton (2015) and Ritchie et al. (2022), the data were then split by performance in the pair (high performer, low performer) to test whether the pairs advantage was driven by improvements in the lower performer. Participants were defined as high and low performers based on their accuracy (percentage correct) at T1. For one pair, both participants had the same accuracy at T1, and so were labelled as high and low performers based on sensitivity (d') measures. Data were entered into a 2 × 3 mixed ANOVA with the factors Pair Member (high, low; between-subjects) and Experimental Session (T1, T2, T3; within-subjects), with the results illustrated in Figure 3. There was a significant main effect of Pair Member, $F(1, 92) = 11.64$, $p = .001$, $\eta_p^2 = .11$, $BF_{10} > 100$, with greater accuracy seen for the high performers ($M = 61.7\%$) compared to the low performers ($M = 56.7\%$). There was also a significant main effect of Experimental Session, $F(2, 184) = 7.61$, $p = .001$, $\eta_p^2 = .08$, $BF_{10} > 100$. This main effect was driven by higher accuracy in T2 compared to T1 ($p = .003$) and T3 ($p = .008$). However, these main effects were qualified by a significant Pair Member × Experimental Session interaction, $F(2, 184) = 11.19$, $p < .001$, $\eta_p^2 = .11$, $BF_{10} > 100$. In order to break down the interaction and test our hypothesis that, as in Ritchie et al. (2022), pairs training would predominantly affect the performance of the low performer in each pair, follow up one-way ANOVAs were conducted for the high and low performers separately.

Accuracy data for the high performers were entered into a one-way repeated measures ANOVA with the factor Experimental Session (T1, T2, and T3). There was no main effect of Experimental Session, $F(2, 92) = 1.70$, $p = .188$, $\eta_p^2 = .04$, $BF_{10} = 0.29$, suggesting high performers remained consistently accurate throughout the three sessions. Data from the low performers were analysed in the same way and showed a significant main effect of Experimental Session, $F(2, 92) = 18.93$, $p < .001$, $\eta_p^2 = .29$, $BF_{10} > 100$. Bonferroni-adjusted pairwise comparisons showed that accuracy at T2 was significantly higher than accuracy at T1 ($p < .001$) and T3 ($p = .003$). In addition, accuracy at T1 was significantly lower than T3 ($p = .017$). This demonstrates a pairs training effect for the low performers, whereby they improved when working in a pair, and also retained some of this benefit when working on their own after this paired interaction.

3.2.2 | Are low performers trained to be as good as high performers?

In line with Ritchie et al. (2022), we tested to see if pairs training improved the low performers to be as good as the high performers. Accuracy data were entered into a 2 × 2 mixed ANOVA with the factors Pair Member (high, low; between-subjects) and Experimental Session (T1, T3; within-subjects). There was no main effect of Experimental Session, $F(1, 92) = .129$, $p = .720$, $\eta_p^2 = .00$, $BF_{10} = 3.00$, and a significant main effect of Pair Member, $F(1, 92) = 22.09$, $p < .001$, $\eta_p^2 = .19$, $BF_{10} > 100$. There was also a significant Pair Member × Time interaction, $F(1, 92) = 11.30$, $p < .001$, $\eta_p^2 = .11$, $BF_{10} = 14.80$. Bonferroni corrected simple main effects analyses demonstrated that the interaction was driven by a significant difference between pair members at T1 ($p < .001$, $d = 1.36$, $BF_{10} > 100$), but not at T3 ($p = .077$, $d = 0.37$, $BF_{10} = 0.88$). This suggests that after working in a pair, the difference

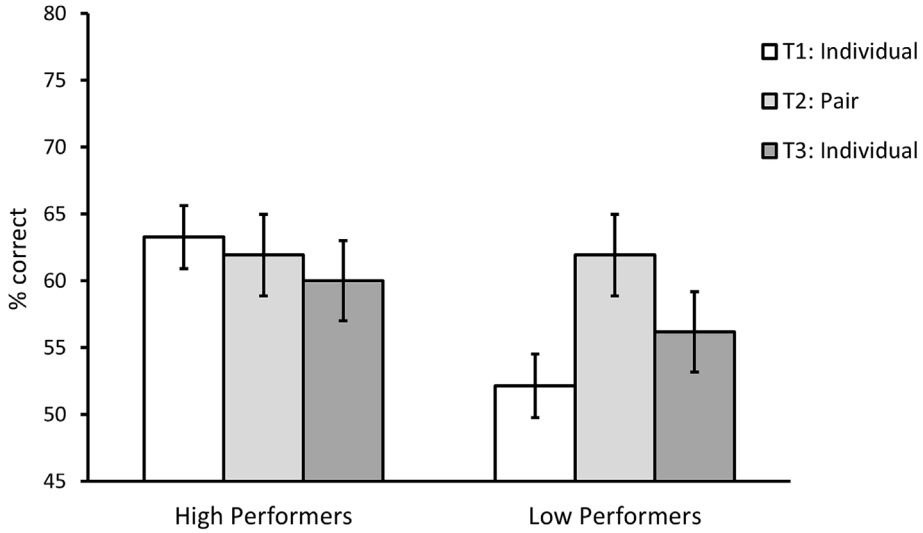

FIGURE 3 Accuracy data for the morph detection task split by performance at T1. Error bars represent 95% confidence intervals of the mean.

in performance between the pair members was no longer present, although caution should be taken as the Bayes Factor for T1 versus T3 suggests only anecdotal evidence for the null hypothesis.

We also explored whether the improvement made by the low performer could be explained by how competent the person was that they worked with. However, this was not found to be the case, as there was a non-significant correlation between the performance of the high performer at T1 and the size of the gain made by the low performer (T3 minus T1), $r(45) = -.09$, $p = .549$, $BF_{10} = 0.22$. We then looked at the difference between the high and low performers at T1, as it is possible that the bigger the initial difference between the pair members, the larger the gain made by the low performer. This correlation was also non-significant, $r(45) = .03$, $p = .051$, $BF_{10} = 1.15$, although this should be interpreted with caution as the Bayes factor indicates anecdotal evidence for the alternative hypothesis.

4 | GENERAL DISCUSSION

The present study investigated whether the pairs training effect, seen previously in unfamiliar face matching, could be demonstrated with face morph detection. This effect has been shown to be a quick and reliable way of improving unfamiliar face matching, and here we demonstrated that working in a pair also improved people's ability to detect face morphs. We were able to show that working in a pair produced better morph detection than working alone, and crucially, that the poorer performer in the pair showed a significant improvement compared to working alone and maintained some of this improvement after working in a pair. There was also some evidence that the lower performers may have learned to be as competent as the higher performers after working in a pair. Interestingly, these improvements were not related to how able the high performer in the pair was, or how much better the high performer was than the lower performer. We ruled out the role of practice effects, differences in block difficulty, and regression to the mean.

Improving face morph detection is becoming increasingly important. Based on work by Kramer et al. (2019), it is unlikely that a morph image would be detected by a human reviewer, and a fraudulently obtained genuine (FOG) document containing the morph image would be mistakenly issued. Once the new passport has been issued, it becomes very difficult to detect the fraudulent document, as it is a genuine and official passport. Previous work has shown mixed evidence for whether participants can improve in morph detection ability, with training unlikely to improve detection of higher-quality morphs (Kramer et al., 2019). With the improvements in image manipulation technology, the majority of face morphs used in the future are likely to be of such high quality that they will be undetectable to the human eye. Modern techniques for creating morphs, such as deep learning-based morph generation, produce very high-quality morphs. These techniques do not need the manual touch-ups that are often required when using standard landmark morph generation (Venkatesh et al., 2021). For the average criminal, landmark techniques are likely to be the most accessible, as software is mainstream and freely available. However, more sophisticated operations are likely to take advantage of these more advanced (and less detectable) techniques. Ultimately, the future solution to the problem is likely to be a computer algorithm. Indeed, Kramer et al. (2019) showed even a simple computer model outperformed their human participants. However wide-scale implementation across government departments requires significant testing, training, and implementation investments. As discussed previously, there are also significant limitations that need to be overcome in order to develop a robust and reliable algorithm appropriate for implementation across official departments (Scherhag et al., 2022). In the meantime, we propose a simple, quick, and reliable form of training that requires minimal effort to implement.

The pairs training effect has previously been shown to be an efficient way to improve unfamiliar face matching. Studies have shown this effect to be reliable and replicable, and the benefit in improvement made by the low performer was maintained even after a delay (Dowsett & Burton, 2015; Ritchie et al., 2022). Given the simple nature of this training, and its applicability to unfamiliar face matching, it is a prime candidate for training in other security-related contexts. Therefore, in this study, we applied the pairs training effect to face morph detection in order to explore whether this form of training could be used to reduce face morphing attacks. Our study supports previous research on the pairs training effect, showing that not only is this effect robust and replicable, but it also transfers to a different type of face task. In addition to being better at morph detection when working in a pair, those who were identified as the lower performers in each pair saw significant improvements in their morph detection accuracy, with some participants seeing improvements of over 20%. For both of the experiments presented here, the high performers' mean accuracy across blocks was 62%, with only one participant scoring above 80% accuracy. This suggests that for high quality morphs, there may be a potential ceiling effect for (human) face morph detection. Although pairs training can significantly improve the lower performers' face morph detection (and bring them in line with the high performers), there is likely a limit to human ability. Our study supports previous research demonstrating that, independent of improvements seen in morph detection rates, face morphs are still a viable method of obtaining and using a FOG document (Kramer et al., 2019; Nightingale et al., 2021; Robertson et al., 2018).

4.1 | Alternative approaches to the problem

Conceptually similar to the current study, 'wisdom of the crowds' has been used to try and improve face morph detection. Responses from multiple participants are aggregated and the majority response is taken as the final answer for each trial. Using this approach, Nightingale et al. (2021) found some improvements in morph detection. However, these improvements were dependant on whether participants had received prior training, as well as the types of mistakes participants were making. For each experiment, the researchers aggregated responses from 100 participants, and as noted by Nightingale et al. (2021), the additional effort of combining responses across multiple individuals might not be feasible in a security context. In contrast, the pairs training effect requires relatively limited effort and resources to implement.

Another approach to the problem of face morph detection (both by humans and algorithms), and potentially the most viable long-term solution, is to have government personnel acquire ID photos (by taking the photographs onsite) at the point of application or place of issue, which prevents individuals from submitting a pre-made morph image. Although this does not prevent other forms of fraud, such as trying to look like someone else through the use of a disguise (Noyes & Jenkins, 2019), it removes the problem of pre-made fraudulent images being submitted by the applicant. Until recently, there has only been limited evidence that this is being considered (Nightingale et al., 2021). However, some governments are now moving towards a 'live capture' process, where ID photos are taken at the passport office (e.g., Huggler, 2020; Immigration Department of Malaysia, 2021) and organisations such as IDEMIA, that specialise in identity-related security, are recommending photo capture by the ID-issuing authority (IDEMIA, 2018). Implementing 'live capture' requires substantial changes to policy and procedures so an interim solution, such as the one presented here, is still required.

4.2 | Further directions

The focus of the present study was to detect morphs at the initial application stage, thereby preventing FOG documents from entering the system and mitigating the risk of their subsequent use in morphing attacks. Inevitably though, FOG documents containing face morph images do end up in use, and it is important to assess their attack potential, at this later stage. A critical direction for future research is the morph attack potential of face morphs created using different methods and software solutions. For an image to be used in a face morph attack, the individual presenting the passport must look sufficiently like the morph for it to be accepted as a 'match'. A current complication for attack detection methods is that individuals presenting themselves at border control can look very different to their passport images (Ferrara et al., 2022; Jenkins et al., 2011). State of the art face recognition systems, such as those used at Automated Border Control (ABC) gates, must be able to tolerate this within-person variability yet remain sensitive enough to reject a morphed image that contains the visual information of another individual. Therefore, a successful face morph must go undetected at the initial application stage *and* fool a face recognition algorithm when presented at an ABC gate.

Another important consideration is the contribution of each individual to the face morph image. In the present study we used 50/50 morphs, equally weighting the two individuals' faces when creating the resulting morph. It is possible to use asymmetric morphing factors, for example 30/70 (Ferrara et al., 2018), or to use different factors for texture blending and shape warping (Ferrara et al., 2019). These different morphing techniques can be used to create morphed faces that are more similar to the accomplice (document applicant) and therefore more difficult to detect at the point of issuance. However, these morphs can have sufficient hidden information of the criminal (document user) that they can still fool a face recognition system (Ferrara et al., 2019). Future studies should consider the *human* detection abilities for these types of morphs and whether this can be improved by the pairs training effect.

5 | CONCLUSION

In the present study, we used high quality face morphs that reflect the quality of morphs likely to be created by a capable fraudster. We showed that simply working together with another person when detecting whether a face image was a morph or not can produce significant improvements in face morph detection. This work has clear implications for real-word security settings, with the goal of providing a quick and reliable way of improving face morph detection.

ACKNOWLEDGEMENTS

We would like to thank Eleanor Earey, Olivia Pine and Tayler Dunn for help with data collection for Experiment 2.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Tessa R. Flack https://orcid.org/0000-0002-4115-4466
Kay L. Ritchie https://orcid.org/0000-0002-1348-760X
Charlotte Cartledge https://orcid.org/0000-0002-0928-0948
Robin S. S. Kramer https://orcid.org/0000-0001-8339-8832

REFERENCES

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207–218. https://doi.org/10.1037/1076-898X.7.3.207

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/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. https://doi.org/10.1111/1467-9280.00144

Dowsett, A. J., & Burton, A. M. (2015). Unfamiliar face matching: Pairs out-perform individuals and provide a route to training. *British Journal of Psychology*, *106*(3), 433–445. https://doi.org/10.1111/bjop.12103

Ferrara, M., & Franco, A. (2022). Morph creation and vulnerability of face recognition systems to morphing. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), *Handbook of digital face manipulation and detection* (pp. 117–137). Springer. https://doi.org/10.1007/978-3-030-87664-7

Ferrara, M., Franco, A., & Maltoni, D. (2014). The magic passport. In *2014 IEEE international joint conference on biometrics* (pp. 1–7). Los Alamitos.

Ferrara, M., Franco, A., & Maltoni, D. (2016). On the effects of image alterations on face recognition accuracy. In T. Bourlai (Ed.), *Face recognition across the imaging spectrum* (pp. 195–222). Springer. https://doi.org/10.1007/978-3-319-28501-6_9

Ferrara, M., Franco, A., & Maltoni, D. (2018). Face demorphing. *IEEE Transactions on Information Forensics and Security*, *13*(4), 1008–1017. https://doi.org/10.1109/TIFS.2017.2777340

Ferrara, M., Franco, A., & Maltoni, D. (2019). Decoupling texture blending and shape warping in face morphing. In *Proceedings of the international conference of the biometrics special interest group (BIOSIG)*.

Ferrara, M., Franco, A., Maltoni, D., & Busch, C. (2022). Morphing attack potential. In *2022 International workshop on biometrics and forensics (IWBF)* (pp. 1–6). Salzburg, Austria. https://doi.org/10.1109/IWBF55382.2022.9794509

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

IDEMIA. (2018). *Facial morphing*. https://epassport-history.idemia.com/downloads/EXT-%20White%20paper%20-%20Facial%20morphing%20-%20April%202019%20-%20ENG.pdf

Immigration Department of Malaysia. (2021). *Malaysian international passport*. https://www.imi.gov.my/index.php/en/main-services/passport/malaysian-international-passport/

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Kaufmann, J., & Schering, A. G. (2007). *Analysis of variance ANOVA*. Wiley Encyclopedia of Clinical Trials. https://doi.org/10.1002/9780471462422.eoct017

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*, 211–222. https://doi.org/10.1002/(SICI)1099-0720(199706)11:3%3C211::AID-ACP430%3E3.0.CO;2-O

Kramer, R. S. S., Mireku, M., Flack, T. R., & Ritchie, K. L. (2019). Face morphing attacks: Investigating detection with humans and computers. *Cognitive Research: Principles and Implications*, *4*(1), 1–15. https://doi.org/10.1186/s41235-019-0181-4

Makrushin, A., Neubert, T., & Dittmann, J. (2017). Automatic generation and detection of visually faultless facial morphs. In F. Imai, A. Tremeau, & J. Braz (Eds.), *Proceedings of the 12th international joint conference on computer vision, imaging and computer graphics theory and applications* (pp. 39–50). Science and Technology Publications. https://doi.org/10.5220/0006131100390050

National Institute of Standards and Technology. (2020). *Face recognition vendor test (FRVT) part 4: MORPH – Performance of automated face morph detection*. https://www.nist.gov/publications/face-recognition-vendor-test-frvt-part-4-morph-performance-automated-face-morph

Neubert, T. (2017). Face morphing detection: An approach based on image degradation analysis. In C. Kraetzer, Y.-Q. Shi, J. Dittmann, & H. J. Kim (Eds.), *Proceedings of the 16th international workshop on digital watermarking* (pp. 93–106). Springer. https://doi.org/10.1007/978-3-319-64185-0_8

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

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

Prolific. (2022). London, UK. https://www.prolific.co/

Qualtrics. (2022). Provo, Utah, USA. https://www.qualtrics.com

Raghavendra, R., Raja, K. B., Venkatesh, S., & Busch, C. (2017a). Face morphing versus face averaging: Vulnerability and detection. In *2017 IEEE international joint conference on biometrics* (pp. 555–563). https://doi.org/10.1109/BTAS.2017.8272742

Raghavendra, R., Raja, K. B., Venkatesh, S., & Busch, C. (2017b). Transferable deep-CNN features for detecting digital and print-scanned morphed face images. In *2017 IEEE conference on computer vision and pattern recognition workshops* (pp. 1822–1830).

Ritchie, K. L., Flack, T. R., Fuller, E. A., Cartledge, C., & Kramer, R. S. S. (2022). The pairs training effect in unfamiliar face matching. *Perception*, *51*(7), 477–495. https://doi.org/10.1177/03010066221096987

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, *141*, 161–169. https://doi.org/10.1016/j.cognition.2015.05.002

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. https://doi.org/10.1371/journal.pone.0173319

Robertson, D. J., Mungall, A., Watson, D. G., Wade, K. A., Nightingale, S. J., & Butler, S. (2018). Detecting morphed passport photos: A training and individual differences approach. *Cognitive Research: Principles and Implications*, *3*(27), 1–11. https://doi.org/10.1186/s41235-018-0113-8

Scherhag, U., Nautsch, A., Rathgeb, C., Gomez-Barrero, M., Veldhuis, R. M. J., Spreeuwers, L., Schils, M., Maltoni, D., Grother, P., Marcel, S., Breithaupt, R., Raghavendra, R., & Busch, C. (2017). Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In *Proceedings of the 2017 international conference of the biometrics special interest group (BIOSIG)*. https://doi.org/10.23919/BIOSIG.2017.8053499

Scherhag, U., Rathgeb, C., & Busch, C. (2022). Face morphing attack detection methods. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), *Handbook of digital face manipulation and detection* (pp. 331–349). Springer. https://doi.org/10.1007/978-3-030-87664-7

Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., & Busch, C. (2019). Face recognition systems under morphing attacks: a survey. *IEEE Access*, *7*, 23012–23026. https://doi.org/10.1109/ACCESS.2019.2899367

Seibold, C., Hilsmann, A., & Eisert, P. (2018). Reflection analysis for face morphing attack detection. In *26th European signal processing conference (EUSIPCO)* (pp. 1022–1026). https://doi.org/10.23919/EUSIPCO.2018.8553116

Spreeuwers, L., Schils, M., Veldhuis, R., & Kelly, U. (2022). Practical evaluation of face morphing attack detection methods. In C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, & C. Busch (Eds.), *Handbook of digital face manipulation and detection* (pp. 331–349). Springer. https://doi.org/10.1007/978-3-030-87664-7

Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, *21*(5), 42–50. https://doi.org/10.1109/38.946630

Venkatesh, S., Ramachandra, R., Raja, K., & Busch, C. (2021). Face morphing attack generation and detection: A comprehensive survey. *IEEE Transactions on Technology and Society*, *2*(3), 128–145. https://doi.org/10.1109/TTS.2021.3066254

White, D., Burton, A. M., Jenkins, R., & Kemp, R. (2014). Redesigning photo-ID to improve unfamiliar face matching performance. *Journal of Experimental Psychology: Applied*, *20*, 166–173. https://doi.org/10.1037/xap0000009

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

How to cite this article: Flack, T. R., Ritchie, K. L., Cartledge, C., Fuller, E. A., & Kramer, R. S. S. (2023). Improving face morph detection with the pairs training effect. *Applied Cognitive Psychology*, *37*(6), 1158–1166. https://doi.org/10.1002/acp.4110