Applied Cognitive Psychology, *Appl. Cognit. Psychol.* **26**: 140–148 (2012)
Published online 31 May 2011 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.1806

In the Dock: Chimeric Image Composites Reduce Identification Accuracy

AILSA STRATHIE[1], ALLAN MCNEILL[1]* and DAVID WHITE[2]

[1]*Department of Psychology, Glasgow Caledonian University, Glasgow, UK*
[2]*Department of Psychology, University of Glasgow, Glasgow, UK*

Summary: The aim of presenting chimeric images (formed from opposing halves of a pair of same or different faces) in court settings is to optimise the accuracy of identification decisions based on CCTV evidence. The experiments reported here examined the utility of this technique. Experiment 1 examined the accuracy of face matching with vertically split, aligned chimeric images, misaligned hemi-faces and full-face images. Experiment 2 replicated the first experiment but replaced the misaligned images with opposing hemi-faces separated by a gap. The final experiment used horizontally split faces. All three experiments showed that matching was less accurate with aligned chimeric images than with full-face images. Furthermore, the pattern of responses obtained with chimeric images differed significantly from full-face matching and misaligned/separated hemi-face matching. Chimeric images produced a bias towards same responses even when the face halves were different. The results suggest caution in the use of chimeric images in court. Copyright © 2011 John Wiley & Sons, Ltd.

INTRODUCTION

Although eyewitness identification evidence can have a powerful effect on a jury, the high value attached to this form of evidence is not always justified, as is evidenced by the alarming number of wrongful convictions attributable to eyewitness misidentifications (Scheck, Neufeld, & Dwyer, 2000). A wealth of psychological research has been conducted with the goal of explaining why these misidentifications occur and addressing the factors that may affect a witness's ability to make an accurate identification (for reviews see Leach, Cutler, & Van Wallendael, 2009; Wells, Memon, & Penrod, 2006; Wells, Small, Penrod, Malpass, Fulero, & Brimacombe, 1998).

Significant progress has been made towards identifying variables that influence accuracy, and research on 'system variables' (Wells, 1978), which are within the control of the legal system, has successfully influenced legal policy and practice (e.g. Wells et al., 2000). Although the police can adopt procedures that minimise the risk of introducing errors, some of the factors that influence the likelihood of an accurate identification, such as the viewing conditions and the quality of a witness's memory, are outside the control of the legal system and their impact can only be estimated after the fact. However, the growing network of closed-circuit television (CCTV) cameras estimated at 60 000 under local authority control in the UK alone (Big Brother Watch, 2009), offers a potential solution to many of these problems.

With a permanent record of a crime captured on CCTV, identification is often reduced to the apparently simpler task of *matching* two faces, one captured at the crime scene with one of the suspect. This alternative method of visual identification is not subject to the memory failures that might undermine eyewitness testimony, and unlike an eyewitness account, CCTV footage is available for visual inspection in court, allowing the jury to make a direct comparison between the two faces for themselves. One might expect that a simple matching task of this type would result in very high levels of accuracy, yet research by Bruce et al. (1999) demonstrated that error rates in unfamiliar face matching tasks are surprisingly high. In their study, experimental witnesses viewed arrays of faces and had to decide whether a simultaneously presented target face was contained in an array of 10 potential matches and, if so, to indicate the person that matched. In target present arrays, participants picked the correct person on around 70% of trials, picked a foil from the array on 12% of trials and indicated the target was not present on the remaining 18% of trials. When the target was absent, participants mistakenly chose someone from the array on roughly 30% of occasions. This poor level of performance is particularly striking because the task used in this study was constructed to optimise performance. There was no memory component, and all images were taken in good lighting from very similar full-face poses. Furthermore, as all images were obtained on the same day, transient differences such as changes in hairstyle were eliminated.

It may be supposed that the high error rates observed in the array task used by Bruce et al. (1999) are partly a function of task difficulty, yet even when task demands are reduced to a verification task between pairs of high quality images, (are both images of the same person?), errors remain high at almost 20% (Burton, White, & McNeill, 2010; Megreya & Burton, 2006, 2007). The important point here is that even in optimal conditions, a simple face matching task is much more difficult than one might imagine. In most criminal cases, the CCTV images are less than optimal, and in such conditions, the task clearly becomes even more challenging. Using a 2 x 4 target present array presentation, Henderson, Bruce and Burton (2001) showed that accurate matching dropped to around 20% when poor quality images (typical of those obtained from commercially available surveillance systems) were used.

Despite the unpromising results obtained in research studies, video identification evidence is well accepted in court (Attorney General's Reference, 2003), and where the images are sufficiently clear, the jury may compare them directly with the defendant in the dock. However, research on live person-to-photo matching suggests that as with photo-to-photo matching, performance is poor (Davis &

*Correspondence to: Dr Allan McNeill, Department of Psychology, Glasgow Caledonian University, Cowcaddens Road, Glasgow G4 0BA, UK.
E-mail: a.mcneill@gcu.ac.uk

Valentine, 2009; Kemp, Towell, & Pike, 1997). In a series of experiments which closely simulate the task a juror might be faced with in court, Davis and Valentine (2009) asked participants to engage in a matching task comparing people who are physically present with simultaneously available video images. Across a series of experiments, these authors consistently found that matching a live person to good quality video footage was highly error prone and was no easier than matching between still images. In their first experiment, participants falsely decided that the target was not present on 22% of target present trials and falsely judged that the images matched on 17% of target absent trials. In subsequent experiments, where the targets were shown in disguise or where the delay between the capture of the 'crime scene' images and the time of the identification was lengthened, accuracy was further reduced.

In a court case, where video quality is high, as in Davis and Valentine's (2009) experiments, jurors are often left to judge for themselves whether the identity of the defendant and the perpetrator match. Where the video footage is less clear, the courts allow for an expert with 'facial mapping' skills to make comparisons between images from the crime scene and images of the defendant and to provide opinion evidence of identity based on the results of these comparisons (Attorney General's Reference, 2003). The Court of Appeal ratified the decision to allow expert testimony on photographic facial image comparison in the courts of England & Wales in 1993 (*R v Stockwell*, 1993), and a few years later, the admissibility of this form of evidence was also confirmed within the Scottish legal system (*Church v HMA*, 1995). The use of these techniques is now well established, and expert testimony based on image comparison is common in cases where identification from photographic material is disputed. Indeed, acceptance of this form of evidence is so complete that convictions have been obtained on the weight of facial image comparison evidence alone (e.g. *R v Hookway*, 1999, *R v Mitchell*, 2005).

The training and background of the experts engaged to conduct facial image comparisons is varied, and the techniques employed are not clearly defined (ACPO, 2003). As a result, the procedures employed vary widely depending on the particular expert engaged but can generally be divided into two categories. The first type focuses on morphological or anthropometric comparisons (comparing features and measuring distance ratios between chosen facial landmarks). The second category involves the combination or superimposition of two images to aid comparison.

Although widely used in court, scientific assessment of the effectiveness of facial image comparison techniques is sparse and has tended to focus on anthropometry (e.g. Davis, Valentine, & Davis, 2010; Kleinberg, Vanezis, & Burton, 2007). Using the 1-in-10 task matching task devised by Bruce et al. (1999), Kleinberg et al. (2007) showed that anthropometric comparisons between target and array faces produced a correct match less than 25% of the time. When compared with human performance on the same task (accurate matching of around 75%), the shortcomings of this technique are clear. Davis et al. (2010) reached a similarly negative conclusion using a computer-assisted decision process. In light of these findings, it is difficult to see the value in continuing to utilise this technique for the purposes of identification.

An illustrative example of the use of techniques that involve the combination of two images for comparison can be seen in the case brought against the Metropolitan Police for a breach of the Health and Safety at Work etc. Act (1974), which it was alleged contributed to the tragic death of Jean Charles de Menezes. In this case, a photographic facial composite formed part of the evidence for the defence. Jean Charles de Menezes was shot after police mistook him for a terrorist suspect who was sought by the police in connection with the failed 21 July London bombings. The defence argued that the wanted man, Hussein Osman, and Jean Charles de Menezes were easily confusable as they were similar in appearance. This argument was supported by a composite image, which consisted of one half of Jean Charles de Menezes's face, combined with the opposite face-half from a photo of Hussein Osman, the man for whom police mistook de Menezes. Although the two men bear some superficial similarities, when the two full-face photographs are examined, it seems unlikely they would be confused. However, when the images are combined to form the composite face, the effect is striking, and subjectively, there is a strong impression that the photo shows two halves of the same face (see Figure 1).

The evidence presented by the defence in the de Menezes case is a little unusual in that it seeks to highlight the similarities between the faces of two different people rather than to establish an identity match. However, in doing so, it powerfully illustrates the dangers of the composite technique and serves to demonstrate how effectively it can convince us that two different people look the same.

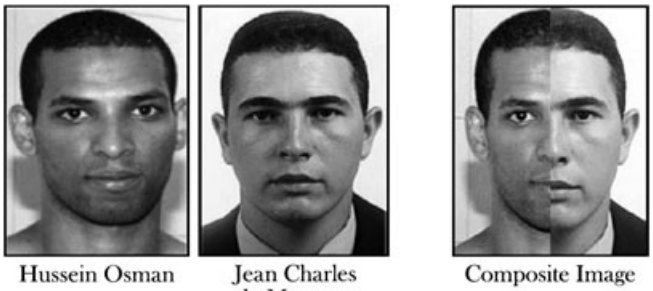

Figure 1. Photos of Hussein Osman, Jean Charles de Menezes and the composite image formed from these two photographs. Source: Metropolitan Police

Unlike anthropometric comparison (Davis et al., 2010; Kleinberg et al., 2007), the practice of using composite photos with the aim of helping people make unfamiliar face matching decisions has not yet been subject to empirical investigation. However, it sits awkwardly with existing research literature on face perception using chimeric images. Using a recognition paradigm with famous faces, Young, Hellawell, and Hay (1987) asked people to look at chimeric faces formed from the top half of one face image and the bottom half of another and to name the person shown in one half of the face. When the two face halves were aligned, people were slower to name the person in the target half of the photo than when the halves were misaligned, or when the composites were inverted (a manipulation, which is believed to interrupt configurational processing). Hole (1994) adopted a matching paradigm to demonstrate that a similar effect also occurs with unfamiliar faces. Participants viewed a series of chimeric face pairs and had to decide whether the top halves of two chimeric faces matched when each was paired with a different bottom half. He found that participants were slower to respond to faces in an upright orientation than to inverted faces. The results of each of these studies indicate that chimeric face images slow responses in tasks that require identification or matching of just one half of a face. Young et al. (1987) and Hole (1994) appeal to the same mechanism to explain this effect and suggest that when two face halves are combined and presented in upright format, they fuse to form a novel face, making it difficult to perceptually separate the two halves into their constituent identities. If the formation of this facial 'gestalt' does underpin these effects, then it is reasonable to predict that the same mechanism will bias observers to say that two *different* faces are the same when opposing hemi-faces are presented in a chimeric format. The premise for the use of such techniques in court is that they improve the accuracy of identification decisions, so our prediction is at odds with established forensic practice.

Here, we present three studies that empirically test this prediction. As the use of chimeric images in facial image comparison techniques is often crucial to the jury's decision, it is forensically important to test this prediction directly by replicating the standard format of these composites. In addition, previous demonstrations of the composite effect have either tested recognition memory (e.g. Young et al., 1987) or have used identical images for matching decisions (e.g. Hole, 1994). In this study, we replicate the applied situation more closely by asking people to compare the two halves of a single chimeric face and by using two different images of the same person in 'same' trials. Experiments 1 and 2 address the principal aim of the study by examining how accurately vertically split chimeric hemi-face images of the type used in the de Menezes case can be matched. Experiment 3 examines matching accuracy with horizontally split images.

EXPERIMENT 1

In this experiment, participants are asked to decide if face pairs show the same person or two different people. The face pairs are presented as full faces, as misaligned opposing hemi-faces, or as chimeric faces formed from opposing hemi-faces—the manipulation employed in the de Menezes case (see Figure 1). These stimulus conditions allow a direct comparison between unfamiliar face matching performance under 'optimal' full-face conditions and performance when matching decisions are made on the basis of chimeric images. We are primarily interested in the difference in responses between the full-face and chimeric face conditions, and we predict more accurate responding for full-face presentations. This prediction is based on the following: (i) the larger amount of information that is available in full-face images; and (ii) the theory (proposed by Young et al., 1987 and Hole, 1994) that chimeric faces appear to form a perceptual whole, thus making it more likely that participants will respond 'same' even for different pairings. The misaligned faces have been included as a control for the chimeric presentation, but making a clear prediction based on previous findings is more problematic. This is because one face-half in the Young et al. (1987) and Hole (1994) studies is task irrelevant and can be easily ignored when the images are misaligned, resulting in less interference compared with chimeric presentations. In the current study, both face-halves are relevant to the task, so ignoring one half-face is not an option. However, one might expect that performance for the misaligned faces will be poorer than for full faces (because only half of the information is available) but better than for chimeric faces (because a misleading facial 'gestalt' is not created).

Method

Participants

Twenty-four students at Glasgow Caledonian University participated in the experiment in exchange for course credit. There were 7 men and 17 women, aged between 17 and 46, [$M = 21.5$, $SD = 7.2$], and all had normal or corrected-to-normal vision.

Design

A 2 x 3 within participants design was utilised. The first independent variable was trial type with two levels, same or different. The second independent variable was presentation format with three levels, Full (two full faces), Misalign (opposing hemi-faces, misaligned) and Chimeric (opposing hemi-faces, aligned). The dependent variable was the accuracy of the same/different decisions, which was measured using sensitivity (d') and criterion (C) values.

Materials

One hundred and fifty face pairs were created using stimuli from the UK Home Office Police Information Technology Organisation database, which contains images of one hundred and twenty trainee policemen aged between 18 and 35 (see Bruce et al., 1999 for full details). Two face pairs were created for each of the targets selected from the database: a matching pair, consisting of two images of the target person (taken using different cameras) presented side-by-side, and a mismatched pair, where the target image was paired with an image of a different person that a group of undergraduate

students had judged to be of similar appearance. These similarity ratings had been previously obtained using a card-sorting procedure outlined by Bruce et al. (1999). Importantly, for matching pairs, the two images used replicate the best conditions for unfamiliar face matching that can be expected in an applied setting when matching a 'mugshot' to a CCTV image. Both images were high resolution and were obtained on the same day under very similar lighting conditions.

All faces were cropped neatly around the contour of the head and were resized so that image height from the top of the head to the tip of the chin was standardised across images. The proportions of the face were held constant. Stimuli for the Full condition were constructed by placing the two whole faces (one from each camera) side-by-side with a small gap of 1 cm between them. Hemi-faces were created by removing pixel information either to the right or left of the axis of symmetry in each full face image pair. In the Chimeric condition, the stimuli were created by fusing together the two half-face images from an opposing hemi-face pair. In the Misalign condition, these two face halves were presented adjacently but were vertically misaligned by one-third of the height of the face. Each of the original image pairs was processed in this manner, and this produced a complete set of 150 pairs (75 same/75 different) in each of the three conditions: 150 full-face pairs, 150 opposing hemi-face misaligned pairs and 150 opposing hemi-face chimeric pairs. This allowed for the items to be fully counterbalanced. All of the images were presented in greyscale. Aside from these manipulations, the images were not altered in any way (e.g. luminance was not normalised) because the aim of the experiment was to replicate as accurately as possible the conditions that could be expected in court, where postproduction is seen as forensic malpractice. Examples of stimuli from each of the conditions can be seen in Figure 2. The stimuli sets were counterbalanced so that across participants, each face pair appeared in each condition an equal number of times.

Procedure

The experiment was administered via a Macintosh laptop running the experimental software PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). Participants were told that on some trials, they would be shown two full faces, whereas other trials would consist of pairs of 'half-faces'. Regardless of the manner in which images were presented, participants were instructed to decide whether the image on the left of the screen showed the same person as the image on the right of the screen and to respond via a keypress. Before the main experiment commenced, participants completed a practice block that contained trials in each of the three presentation formats to ensure they fully understood the task. Each trial commenced with a fixation cross displayed on screen for 1 second followed by one of the image pairs. The image pair remained on screen until participants made their decision. Each participant viewed 150 trials, 50 (25 same and 25 different) trials in each condition, intermixed in random order. There were no breaks.

Results

Sensitivity (d') and criterion (C) scores were calculated for each participant at each level of presentation format. Table 1 shows the mean values with standard deviations for each condition.

The sensitivity (d') and criterion (C) scores were analysed using two separate one-way ANOVAs. For sensitivity, a significant main effect of presentation format was found, $F(2,46) = 87.63$, $p < .05$. Tukey HSD tests (at $p < .05$) revealed that sensitivity in the Full condition was significantly higher than sensitivity in both the Misalign and Chimeric conditions. Sensitivity in the Misalign condition was also significantly higher than in the Chimeric condition. All sensitivity scores were significantly above chance performance (hypothetical mean = 0, $p < .05$).

For criterion, a significant main effect of presentation format was also found, $F(2,46) = 11.96$, $p < .05$. Tukey HSD tests (at $p < .05$) revealed a significant difference in response bias between the Full and Chimeric conditions and between the Misalign and Chimeric conditions. The difference between Full and Misalign conditions was not significant. The positive criterion value in the Misalign condition suggests that participants exhibit a bias towards 'different' responses in this condition. In contrast, the negative value obtained in the Chimeric condition suggests that participants

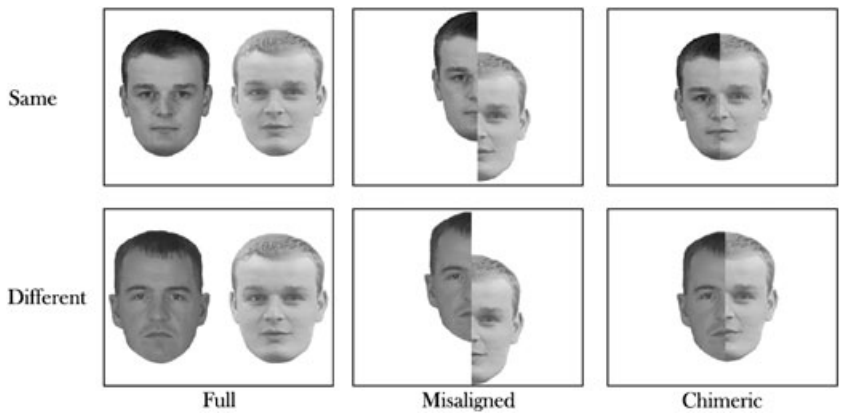

Figure 2. Examples of stimuli from each experimental condition in experiment 1

Table 1. Mean signal detection measures with standard deviations for experiment 1

	Presentation format	Mean	*SD*
Sensitivity (*d'*)	Full	2.92	0.85
	Misalign	1.44	0.53
	Chimeric	1.05	0.65
Criterion (*C*)	Full	0.05	0.58
	Misalign	0.27	0.45
	Chimeric	−0.26	0.47

are biased towards responding 'same' in this condition. To test this assertion, we carried out three separate one sample *t*-tests comparing criterion scores against a chance score of zero (i.e. no bias). No response bias was observed in the Full condition, $t(23) = 0.45$, $p > .05$, however, significant bias was detected for both Misalign, $t(23) = 2.88$, $p < .05$ and Chimeric, $t(23) = -2.75$, $p < .05$ conditions.

Discussion

As predicted, the main effect of presentation format (Chimeric/Misalign/Full) demonstrates participants performed best with full-face images (88% correct), followed by the Misalign condition (73% correct), followed by the Chimeric condition (67% correct). The results are better understood when performance is broken down into sensitivity and criterion components.

As expected, sensitivity was superior for full face stimuli than for both types of hemi-face presentation. Although we predicted this result, the outcome was not a certainty as familiar face recognition is highly robust to large distortions and is often unaffected by severe image degradation (e.g. Hole, George, Eaves, & Razek, 2002). However, the results of the current study demonstrate that when the facial stimuli are unfamiliar to participants, these manipulations result in a marked decrement in performance. Furthermore, when hemi-face stimuli were presented in misalignment, performance was improved relative to when the face halves were fully aligned. This manipulation also had a significant effect on participants' response bias. Specifically, when hemi-faces were presented as chimeric images (as they often are in court), there was a significant bias towards responding that the two images were of the same person. In contrast, misalignment of the face halves induces more 'different' responses. These results show for the first time that chimeric image presentations of the type used by facial mapping practitioners tend to impede rather than optimise the accuracy of unfamiliar face matching decisions. As images of this type are regularly used in courtroom settings, the implications of these findings are of considerable applied importance.

EXPERIMENT 2

In experiment 1, we demonstrated that matching identities from misaligned hemi-faces or chimeric faces is less accurate than matching with full-face images. Furthermore, when two face halves are fused together to form chimeric faces, this results in a bias towards 'same' responses. Misaligning the hemi-faces produces the reverse pattern of performance, evoking a bias towards 'different' responses. In this experiment, we seek to replicate the advantage for full faces in terms of sensitivity, and also the bias towards responding 'same' to chimeric faces. Further, we hope to clarify the basis of the bias towards 'different' responses to same faces observed in the misaligned condition. If responding same to 'different' faces in the chimeric condition is underpinned by the formation of a facial 'gestalt' that contains illusory continuities (the apparent lining up of eyes, ears and mouth for example), then it may be that misaligned images of the same faces (in which pre-existing natural continuities are destroyed) create a bias towards different responses. Therefore, in experiment 2, we will replace the misaligned faces from experiment 1 with faces that are correctly aligned but have a gap between the two halves. If the alignment of the faces were responsible for the bias towards more conservative responding with misaligned faces in experiment 1, then we would expect this bias to be reduced in the 'gap' condition of this experiment.

Method

Participants

Twenty-four students at Glasgow Caledonian University participated in the experiment in exchange for course credit. There were six men and 18 women, aged between 18 and 29 [$M = 19.6$, $SD = 4.0$], and all had normal or corrected-to-normal vision.

Design and procedure

The design and procedure were the same as for experiment 1, save that the Misalign condition was replaced by the Gap condition.

Materials

The materials for the full face and Chimeric conditions were the same as in experiment 1, except that in place of the Misaligned condition, two face halves were presented correctly aligned but with a small gap in between. As in experiment 1, this produced a set of 150 pairs (75 same/75 different) in each of the three conditions: 150 full-face pairs, 150 half-face pairs with a gap of a face width and 150 half-face chimeric pairs. All stimuli were presented in greyscale. Examples of stimuli from each of the conditions can be seen in Figure 3.

Results

Sensitivity (*d'*) and criterion (*C*) scores were calculated for each participant at each level of presentation format. Table 2 shows the mean values with standard deviations for each condition.

Two one-way within-subjects ANOVAs were conducted to examine sensitivity (*d'*) and criterion (*C*) scores. For sensitivity, a significant main effect of presentation format was found, $F(2,46) = 44.05$, $p < .05$. Tukey HSD tests (at $p < .05$) revealed that sensitivity was significantly higher in the Full condition than in both the Central Gap and Chimeric conditions. The difference in sensitivity between Gap and Chimeric conditions was not significant. Sensitivity scores for all three presentation formats significantly exceeded chance (hypothetical mean = 0, $p < .05$). A significant main

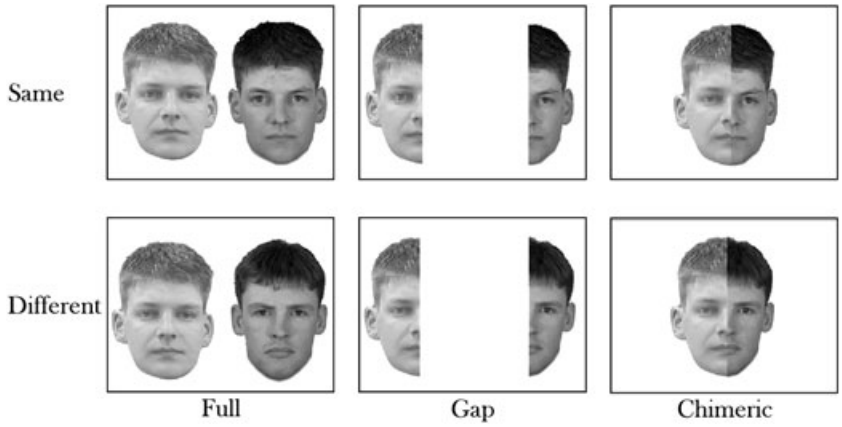

Figure 3. Examples of stimuli from each experimental condition in experiment 2

Table 2. Mean signal detection measures with standard deviations for experiment 2

	Presentation format	Mean	*SD*
Sensitivity (*d'*)	Full	2.70	0.59
	Gap	1.52	0.47
	Chimeric	1.33	0.68
Criterion (*C*)	Full	–0.12	0.51
	Gap	0.36	0.45
	Chimeric	–0.47	0.46

effect of presentation format was also found for criterion, $F(2,46) = 32.97$, $p < .05$. Tukey HSD tests (at $p < .05$) revealed a significant difference in response bias between Full and Gap, Full and Chimeric and Chimeric and Gap conditions. The positive criterion value in the Gap condition suggests that participants exhibit a bias towards 'different' responses in this condition. In contrast, the negative value obtained in the Chimeric condition suggests that participants are biased towards responding 'same' in this condition. To confirm that these biases differed from baseline, we again carried out three separate one-sample *t*-tests comparing criterion scores against a chance score of zero (i.e. no bias). As with experiment 1, no response bias was observed in the Full condition, $t(23) = -1.18$, $p > .05$, however, there were significant biases in both Gap, $t(23) = 3.87$, $p < .05$ and Chimeric conditions, $t(23) = -4.98$, $p < .05$.

Discussion

In this task, participants performed best in the Full image condition (88% correct), with mean accuracy of 90% on 'same' trials and 86% on 'different' trials. The significant effect of presentation format (Chimeric/Gap/Full) on sensitivity demonstrates that as predicted, performance is poorer in both half-face conditions (Gap, 74% correct; Chimeric, 70% correct) relative to full-face matching, thus replicating the full-face advantage seen in experiment 1. Although there was no significant difference between Gap and Chimeric conditions for sensitivity, when criterion is examined, it is clear that each of the hemi-face manipulations has a different effect on response bias; Chimeric presentations encourage a bias towards 'same' responses, whereas Gap trials biased responses in the opposite direction.

Importantly, this replicates the finding observed in experiment 1 and provides further evidence that the chimeric presentation of faces tends to reduce, rather than improve, the accuracy of unfamiliar face matching decisions. The tendency towards responding 'different' on same trials in the Gap condition suggests that separating the hemi-faces in space evokes a similar bias to that observed with the misaligned face halves in experiment 1. One interpretation of this finding is that the decrease in accuracy with hemi-face presentations simply results from the reduction in information available in the half faces relative to full-face presentations. With less information to inform their decisions, participants may, in general, be more conservative when making these judgements. However, relative to full faces, there is also less information in the chimeric images, yet a liberal response bias is observed in this condition. This may suggest that the chimeric presentations evoke holistic processing, creating the perception of a whole face, whereas either misaligning or separating the faces in space interrupts this processing, emphasising the differences between the two hemi-faces and encouraging 'different' responses. Relative to the misaligned presentation in experiment 1, the gap presentation in the current experiment did not appear to reduce the tendency towards a conservative bias, though this may be a result of the size of the gap used here. Reducing the size of the gap may moderate this effect.

EXPERIMENT 3

Having established in experiments 1 and 2 that matching decisions based on composite images formed from vertically split hemi-faces results in a tendency towards 'same' responses, we sought to determine whether the same bias is apparent with horizontally split faces. In this condition, there is very little information in the top half of the face that can predict the appearance of the bottom half of the face, so we expected that overall performance would be poorer than with vertically split images.

Method

Materials

The full-face images were the same as those used in experiments 1 and 2. To create the stimuli for the two experimental conditions, the full face images were modified

as follows: Faces were divided horizontally along the middle of the nose, and pixel information was removed from opposing sides of this dividing line from each of the two images in the full face pair. These half face images were then aligned and were manipulated to form a pair consisting of two opposite face halves separated by a gap of one face height, (Horizontal Gap condition) and a pair where the two face halves were fused together (Chimeric condition). This process was repeated for all 150 full face-pairs to produce a set of 150 pairs (75 same/75 different) in each of the three conditions: 150 full-face pairs, 150 half-face with central gap pairs and 150 half-face chimeric pairs. All stimuli were presented in greyscale. Examples of stimuli from each of the conditions can be seen in Figure 4.

Participants

Twenty-four students at Glasgow Caledonian University participated in the experiment in exchange for course credit. There were three men and 21 women aged between 17 and 40 [$M = 20$, $SD = 5.3$], and all had normal, or corrected-to-normal, vision.

Design and procedure

The design and procedure were the same as for experiment 2 except that in Gap and Chimeric conditions, the images were split horizontally rather than vertically.

Results

Following the same procedure as in the previous experiments, sensitivity (d') and criterion (C) scores were calculated for each participant (see Table 3) and analysed using two one-way within-subjects ANOVAs.

For sensitivity, a significant main effect of presentation format was found, $F(2,46) = 140.78$, $p < .05$. Tukey HSD tests (at $p < .05$) revealed that sensitivity was significantly lower in the Chimeric and Gap conditions than in the Full Face condition. The difference between Gap and Chimeric conditions did not reach significance. When sensitivity scores were compared with chance performance (zero), we found that although both full and misaligned conditions were above chance ($p < .05$), sensitivity in the chimeric condition did not differ significantly from chance, $t(23) = 0.97$, $p > .05$.

Table 3. Mean signal detection measures with standard deviations for experiment 3

	Presentation format	Mean	*SD*
Sensitivity (d')	Full	2.70	0.86
	Gap	0.28	0.31
	Chimeric	0.09	0.45
Criterion (C)	Full	−0.13	0.53
	Gap	0.35	0.28
	Chimeric	−0.14	0.40

For criterion, a significant main effect of presentation format was also found, $F(2,46) = 9.16$, $p < .05$. Tukey HSD tests (at $p < .05$) revealed a significant difference in response bias between the Full and Gap conditions and between the Gap and Chimeric conditions. The difference between Full and Chimeric conditions was not significant. Criterion scores were tested against a hypothetical mean of zero, using three separate one-sample t-tests. As with both previous experiments, no significant bias was detected in the Full condition, $t(23) = -1.22$, $p > .05$. Although a bias towards 'different' responses was observed in the Gap condition, $t(23) = 6.1$, $p < .05$, the tendency to respond 'same' in the Chimeric condition was not significantly different from zero in this experiment, $t(23) = -1.8$, $p > .05$.

Discussion

In this experiment, participants performed better with Full face images (87% correct) than in the Gap condition (55% correct) or the Chimeric condition (52% correct). The sensitivity of matching decisions based on horizontally split hemi-faces was not above chance level when images were presented as a chimera and was very slightly above chance when hemi-faces were presented with an intervening gap. This indicates that information in the top half of a face cannot be used to reliably predict the appearance of the bottom half of the face. The decrement in hemi-face

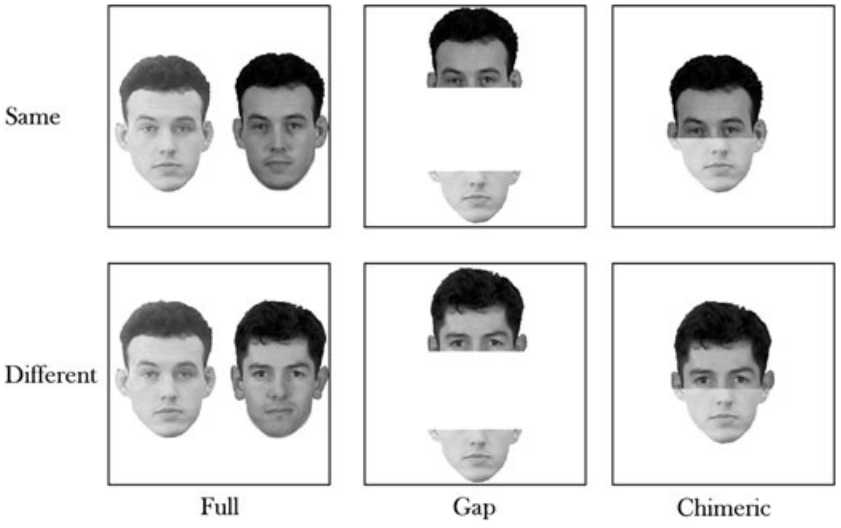

Figure 4. Examples of stimuli from each experimental condition in experiment 3

matching accuracy in this experiment relative to that observed with vertically split experiments is not unexpected. A likely explanation for this finding is that although certain visual information is correlated across two hemi-faces (i.e. our two eyes share more similarities than differences; see Farkas & Cheung, 1981), faces are clearly not symmetrical with respect to the horizontal axis. Despite the poor performance on this task, participants showed the same pattern of responses identified in experiments 1 and 2 with gap presentations evoking a higher proportion of 'different' responses. However, although response bias in the Chimeric condition was significantly different from response bias in the Gap condition, it was not significantly different to baseline criterion. Therefore the pattern observed in criterion scores for experiments 1 and 2 was only partially replicated.

GENERAL DISCUSSION

Data from these experiments replicates previous demonstrations of generally poor performance in 1:1 unfamiliar face matching. Across three experiments, we found that under optimal matching conditions (i.e. Full face condition) in a two alternative forced choice task, participants made an average of 13% errors. This poor level of performance corroborates previous research that demonstrates that unfamiliar face matching based on photographic evidence is inherently unreliable (e.g. Bruce et al., 1999; Burton et al., 2010; Kemp et al., 1997). Furthermore, we found that the poor performance observed with this task was exacerbated by the use of chimeric stimuli.

In experiment 1, we established that the use of composite images did not increase the accuracy of unfamiliar face matching decisions, and, as hypothesised, accuracy was better with full faces than with either of the composite presentations. Importantly, the use of a 'chimeric' image of the type used in the de Menezes case resulted in an increase in false positive responses to images consisting of two different faces. Interestingly, we observed the opposite pattern of errors when the hemi-faces were misaligned, with these presentations producing a bias towards false negative responses for same trials.

Experiment 2 replicates the main findings from experiment 1. Overall performance was significantly poorer in both the composite face conditions compared with the full face condition. Once more, the use of chimeric type composite images appeared to increase the likelihood of participants' falsely categorising images of two different faces as a match. Also consistent with experiment 1 is the finding that disrupting the alignment of chimeric stimuli (in this case by introducing a gap) evokes a more conservative response strategy. In future work, it might be interesting to investigate whether the response bias in gap conditions varies as a function of the degree of separation, or whether any gap, regardless of size, is sufficient to induce the positive criterion values observed here. A recent study by Taubert and Alais (2009), which examined the effects of varying the gap size with horizontally split composites, suggests a gap of one-half a face height is sufficient to disrupt holistic processing, whereas a gap of one-fourth of a face height is not. They suggest that this is because the composite face must appear biologically plausible to evoke the composite face effect. This explanation would predict that a larger gap, of the size employed in the current experiment and in experiment 3, makes the face appear biologically implausible, and this may explain the bias towards 'different' responses with this presentation mode.

Experiment 3 demonstrated that the effects observed with vertically split hemi-faces also operate when these presentation modes are applied to horizontally split hemi-faces. However, the response bias for chimeric stimuli did not differ from chance in this condition. This may suggest that the holistic mechanisms responsible for the increase in same responses in the previous two experiments were weaker for horizontally split than vertically split stimuli. This is a novel finding and might be explained by the relative importance of mirror symmetry as an organising factor in visual perception (e.g. Barlow & Reeves, 1979; Koffka, 1935). We find this to be an interesting possibility, which could be investigated in future research.

Taken together, our findings are consistent with current theories of face processing, which suggest that people generally use some form of 'holistic' processing, in which the whole face is processed without explicit recognition of the constituent parts (e.g. Hole, 1994; Tanaka & Farah, 1993; Tanaka & Sengco, 1997; Young et al., 1987) rather than using a purely feature-based approach. In line with Young et al. (1987) and Hole (1994), we suggest that the use of chimeric images encourages holistic type processing of the resultant face. A consequence of this appears to be a bias towards saying that the halves of two different faces belong to the same person when they are fully aligned. The bias towards different responses for same face-halves when they are presented misaligned or with a gap might also be explained by appealing to the same framework. In this situation, the constituent parts are already isolated in space. We have little experience of viewing half faces of this type, so it seems reasonable to speculate that the usual holistic face processing mechanisms will not be engaged with stimuli of this type. In such circumstances, it is probable that participants will resort to using a feature-based processing strategy. Additionally, our lack of experience with matching face halves of this type may encourage caution, promoting a more conservative response strategy. Alternatively, the tendency towards different responses may result from the biological implausibility (Taubert & Alais, 2009) of these presentations. In future research, this question could be investigated further by presenting stimuli upside down to assess how disrupting these configural processing mechanisms affects responses.

The results of these experiments, taken as a whole, suggest that showing a jury composite images formed from two different photographs is unlikely to make a useful contribution to the process of identity matching. These images not only reduce the overall accuracy of matching decisions relative to full face matching but also increase the likelihood of a 'same' response to a chimeric image formed from two different faces. However, it should be noted that the experimental manipulations employed in the current study do not directly simulate the typical use of these techniques in a

forensic setting. In a court case, the original photographic material from which the composite image was formed would normally be available to the jury, and, in most cases, the defendant would be physically present. It is possible that these factors could offer the jury some protection from the biasing effect of superimposed or composite photographs. However, the documented use of the composite image created from images of Jean Charles de Menezes and Hussein Osman demonstrates that chimeric images may sometimes be presented under similar conditions to those employed in the current study. This practice is of particular concern given the outcome of the experiments reported here.

A further point of departure from forensic practice is that in a court setting, the jury would normally be asked to make just one matching decision rather than a series. Therefore, they are likely to spend more time examining the images, and an awareness of the importance of their conclusion may motivate jurors to take greater care over this decision. The impact of these factors warrants investigation, but even if they do attenuate the biasing effect of composite face images, it remains unlikely that a composite image will produce an *increase* in accuracy. A further consideration is that when composite photographs are presented in court, the visual evidence is supported by testimony from an expert witness. The expert is likely to highlight facial similarities observed in the two images and may also present a statement regarding the degree of support found to suggest that the images are of the same person. It is possible that in conjunction with expert evidence, the effects of composite faces techniques may be even more compelling and in this situation, their misleading influence may be strengthened.

This paper represents the first attempt to objectively evaluate the utility of composite face matching techniques. The data here indicate that these techniques impair rather than improve the accuracy of identity matching decisions. Perhaps most worryingly, the use of composite face images appears to increase the probability that images of two different people will be judged to show the same person, thus increasing the likelihood of wrongful conviction. In conclusion, the data presented here suggest that composite face matching techniques should not be used in a courtroom setting.

REFERENCES

Association of Chief Police Officers (ACPO). (2003). National working practices in facial imaging. http://www.acpo.police.uk/asp/policies/Data/garvin_facial_imaging_guidelines.doc [accessed on 30 July 2008]

Attorney General's Reference. [no.2 of 2002], (2003). 1 CR App R 321, England.

Barlow, H. B., & Reeves, B. C. (1979). Versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, *19*(7), 783–793.

Big Brother Watch. (2009). Big Brother is watching: The first comprehensive analysis of the number of CCTV cameras controlled by local authorities in Britain in 2009. http://www.bigbrotherwatch.org.uk/cctvreport.pdf [accessed on 20 January 2010]

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

Burton, A. M., White, D., & McNeill, A. (2010). The glasgow face matching test. *Behavior Research Methods*, *42*, 286–291.

Church *v* HMA. (1995). SLT 604.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments and Computers*, *25*, 257–271.

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology*, *23*(4), 482–505. DOI: 10.1002/acp.1490

Davis, J. P., Valentine, T., & Davis, R. E. (2010). Computer assisted photo-anthropometric analyses of full-face and profile facial images. *Forensic Science International*, *200*, 165–176.

Farkas, L. G., & Cheung, G. (1981). Facial asymmetry in healthy North American Caucasians—An anthropometrical study. *The Angle Orthodontist*, *51*(1), 70–77.

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, *15*, 445–464. DOI: 10.1002/acp.718

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception 23*, 65–74.

Hole, G. J., George, P. A., Eaves, K., & Razek, A. (2002). Effects of geometric distortions on face recognition performance. *Perception*, *31* (10), 1221–1240.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*, 211–222.

Kleinberg, K. F., Vanezis, P., & Burton, A. M. (2007). Failure of anthropometry as a facial identification technique using high-quality photographs. *Journal of Forensic Sciences*, *52*(4), 779–783.

Koffka, K. (1935). *Principles of gestalt psychology*. New York: Harcourt.

Leach, A., Cutler, B. L., & Van Wallendael, L. R. (2009). Lineups and eyewitness identification. *Annual Review of Law and Social Science*, *5*, 157–17.

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

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

R *v* Hookway. (1999). EWCA Crim 212.

R *v* Mitchell. (2005). EWCA Crim 731.

R *v* Stockwell. (1993). 97 Cr App R 260, England.

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual innocence: Five days to execution and other dispatches from the wrongly convicted*. New York: Doubleday.

Tanaka, J. W., & Farah, M. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*, 225–245.

Tanaka, J. W., & Sengco, J. A. (1997). Features and their configuration in face recognition. *Memory & Cognition*, *25*, 583–592.

Taubert, J., & Alais, D. (2009). The composite illusion requires composite face stimuli to be biologically plausible. *Vision Research*, *49*, 1877–1885.

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Experimental Psychology*, *36*, 1546–1557.

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *The American Psychologist*, *55*(6), 581–598.

Wells, G. L., Memon, A., & Penrod, S. (2006). Eyewitness evidence. Improving its probative value. *Psychological science in the Public Interest*, *7*(2), 45–75.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, *23*(6), 603–647.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*, 747–759.