Running head: Morphing facial composites

Evolving and combining facial composites: Between-witness and within-witness morphs compared.

Tim Valentine[1], Josh P. Davis[2], Kate Thorner[1]

Chris Solomon[3] & Stuart Gibson[3]

[1] Goldsmiths, University of London; [2] University of Greenwich; [3] University of Kent

Keywords: facial, composite, identification, morph, witness

Author Note

We gratefully acknowledge financial support from the EPSRC (Grant number GR/S98511/01). Experiments 2 and 4 were presented to the 25th Conference of the British Psychological Society Cognitive Section, Southampton, September 2008. Chris Solomon and Stuart Gibson are faculty members of the University of Kent and Directors of Vision Metric Ltd. Vision Metric Ltd. markets EFIT-V, a commercially available software package developed from the EigenFIT software used to generate the composites in these experiments. Solomon and Gibson played no part in the collection, analysis or interpretation of data. Their contribution as authors is the intellectual contribution in developing the software and providing software support.

Abstract

Student participant-witnesses produced four composites of unfamiliar faces with a system that uses a genetic algorithm to evolve appearance of artificial faces. Morphs of four composites produced by different witnesses (between-witness morphs) were judged better likenesses (Experiment 1) and were more frequently named (Experiment 2) by student participants, who were familiar with the target actors, than were morphs of four composites produced by a single witness (within-witness morphs). Within-witness morphs were judged better likenesses and more frequently named than the 'best' or the first-produced individual composites. The same results for likeness judgements were observed after possible artefacts in the comparison of between- and within-witness morphs were eliminated (Experiment 3). Experiment 4 showed that both internal and external features were better represented in morphs than in the original composites, although the representation of internal features improved more. The results suggest that morphing improves the representation of faces by reducing random error. Between-witness morphs yield more benefit than within-witness morphs by reducing consistent but idiosyncratic errors of individual witnesses.

The experiments provide the first demonstration of an advantage for within-witness morphs produced using a single system. Experiment 2 provides the first demonstration of a reliable advantage for naming between-witness morphs in the most forensically relevant task: naming a composite of a familiar person produced by a witness who was unfamiliar with the target. Morphing would enhance the recognition of facial composites of criminals. Within-witness morphing provides a methodology for use in many crimes in which the victim is the only witness.

Evolving and combining facial composites: Between-witness and within-witness morphs compared.

In the early stages of a criminal investigation police officers often face the difficult problem of identifying a possible suspect. In the absence of identification from forensic evidence (e.g., a match to a DNA database), police may ask a victim to assist in the production of a facial likeness (or composite) of the offender. The methods used to construct facial composites have developed considerably over the years from a police artist's sketch, through mechanical systems (e.g., Identi-Kit, Photo-Fit), and software versions (e.g., E-FIT, PRO-fit, Mac-a-Mug, FACES), to development of systems that use a genetic algorithm to 'evolve' an artificial facial likeness (e.g., EFIT-V and EvoFIT). Research has demonstrated that: 1) all systems produce facial composites that are rather poorly recognised by people who know the target person; and 2) generally composites produced by a skilled police artist can outperform systems (e.g., Frowd, Carson, Ness, McQuiston, Richardson, Baldwin, & Hancock, 2005a; see Davies & Valentine, 2007 for a review). Excluding circumstances when there is a strong contextual cue, the typical identification rate of composites constructed of an unfamiliar face, from memory, under forensically plausible conditions is less than 4% of people who are familiar with the target face. A police artist's sketches were identified, under the same conditions, by 8% of participants (Frowd *et al*., 2005a). Higher naming rates from composite systems (up to 33%) have been reported under forensically-relevant conditions for target faces selected to be distinctive (Frowd, Carson, Ness, Richardson, Morrison, McLanaghan & Hancock, 2005b).

The disappointing performance of facial composite generators may, in part at least, be attributed to the piecemeal manner in which witnesses are required to select individual facial features in isolation. The psychological science of face recognition shows that this approach goes against the grain of our natural face processing strategy, which exploits our ability to process the configural or holistic properties of a face (e.g., Leder & Bruce, 2000; Tanaka & Farah, 1993). Therefore, assembling a face from a montage of individual features, by selecting the eyes, nose or mouth independently, is a process that we are likely to find extremely difficult. According to this view it is unsurprising that people often fail to create effective facial composites. However, two recent developments are more soundly based in the theory of human face perception: morphing between composites generated by different witnesses; and the use of genetic algorithms to evolve the appearance of artificial faces under the guidance of the witness. These innovations were combined and developed in the experiments reported here.

Witnesses are likely to notice and remember the distinctive features of a face (Light, Kayra-Stuart & Hollander, 1979; Valentine, 1991; Winograd, 1981). Therefore, composites constructed by eyewitnesses are more likely to be consistent in the representation of distinctive features. If witnesses include erroneous features, it is advantageous that the errors should be typical or average in appearance, so that they are less salient in processing the face. If the errors that witnesses make are random, or idiosyncratic, morphing composites made by different witnesses will reduce the distinctiveness of features not shared by all the composites. At the same time the distinctiveness of features common to all the composites will be maintained. Empirical support for this argument was reported by Bruce, Ness, Hancock, Newman and Rarity

(2002). A morph of four facial composites produced by different witnesses of a familiar face and of an unfamiliar face, produced from memory and with a photograph present, was rated as a better likeness of the target than were the individual composites. In a further experiment Bruce *et al.* explored the ability of people who were familiar with the target face to name composites and the morphs of four composites produced by different witnesses who were unfamiliar with the target person. Showing all four individual composites produced the highest rate of naming (38%), the morph was named by 28%, the best individual composite was named by 22% of participants and the worst individual composite was named by 16%. It should be noted that these differences were not statistically significant. The relatively high rate of naming reported occurred in a context in which participants knew the target people were members of staff in a university department, therefore they had a relatively strong contextual cue to naming.

Morphs are more average in appearance than individual composites as a result of averaging across four images. Hasel and Wells (2007) pointed out two possible artefacts. First, morphs are likely to be similar to more faces than is an individual composite because they are averaged (a prototype effect). Therefore it is likely that morphs will be more similar to non-targets as well as to the target. Second, average faces are perceived as more attractive (e.g., Rhodes & Tremawan, 1996, Valentine, Darling & Donnelly, 2004). Therefore, morphing composites may be less effective for less attractive faces (an attractiveness effect). Hasel and Wells found evidence for effects of both artefacts on likeness judgements. Nevertheless, morphs of four composites compiled by four different witnesses were judged to be a better likeness to the target even when the effects of these artefacts were controlled.

The advantage for morphing composites is of practical value because the police have no way of knowing which witness has produced the best composite, or whether self-ratings by different witnesses reliably reflect the quality of the likeness. Therefore, it is useful to demonstrate that a morph is at least as good as the best individual composite. If this is the case the police can use a morph knowing that the image will be at least as effective as any of the individual composites available. However, the benefit of morphing composites produced by different witnesses is limited to cases in which there are multiple witnesses available. In this paper we use a novel holistic method of generating facial composites (EFIT-V[1]) to compare the effectiveness of individual composites and their morphs. EFIT-V is used by 16 police forces in the UK. Only E-FIT is used by more British police forces. EFIT-V allows a likeness to be produced much more quickly than is possible using a conventional composite system, making it possible for each participant to construct a composite of each target four times. Therefore it becomes possible to compare between-witness morphs (as evaluated by Bruce *et al.*, 2002 and Hasel & Wells, 2007) and within-witness morphs, produced by a weighted morph of four composites created by the same witness. If within-witness morphs produce a likeness as good or better than the best individual composite it would allow the advantage of morphing composites to be gained in any investigation that has only a single witness.

When constructing a composite using EFIT-V the witness selects a category of the target face (e.g., white male) and a hairstyle from a menu of alternatives in the usual manner of composite generators. The witness then sees a screen of nine random artificially generated faces that all share the chosen hairstyle. Although none will closely resemble the target, the witness is asked to select the face that most resembles the target.

This selection is then used to generate a new set of nine artificial faces. Each successive generation of evolved faces is generated by use of a genetic algorithm, which applies random variation to the selected facial appearance. The witness selects the face that most resembles the target from each successive screen of nine faces in an iterative process. The facial appearance of the artificial faces will converge on the desired appearance, so that the faces on each successive screen become more similar to each other. The process stops when the witness is satisfied with the appearance of a face or when all the faces are so similar it is not possible to choose between them.

The process of evolving the facial appearance in EFIT-V is analogous to gradient descent of error in a neural network. For the first generation the error (difference in appearance to the target) is high. With each successive choice, the witness specifies the direction to change the 'seed' face for production of the next generation (which includes random variation). Gradually the rate of gradient descent decreases until the face is located at a minimum of error. The starting point (i.e. the generation of the first screen of faces) is random. If a witness attempts to construct a likeness of the same target a second time, she or he will start with different artificial faces and so may finish with a different final composite. EFIT-V has a number of software tools and options to make adjustments to a selected image in a more traditional feature-based manipulation. For example, features can be re-sized or re-positioned, the age of the face can be adjusted, and a feature can be 'locked' so that it does not change in any subsequent generations. For further information on EFIT-V (EigenFIT) see Gibson, Pallares Bejareno and Solomon (2003), Gibson, Solomon, Maylin, and Clark (2009) and Davies and Valentine (2007).

EFIT-V has not been evaluated previously, but a system that works on a similar principle, EvoFIT has been evaluated by Frowd and colleagues (Frowd, Hancock & Carson, 2004; Frowd *et al.* 2005a and 205b; Frowd, Bruce, Ness, Bowie, Paterson, Thomson-Bogner, McIntyre & Hancock, 2007). The early version of Evo-Fit generally performed at a level similar to the traditional software system (e.g., E-FIT). Sketches by a police artist were identified more frequently by people familiar with the target person (8.1%) than were composites produced using EvoFIT (3.6%) or PRO-Fit (1.3%) (Frowd *et al.*, 2005a). Comparative studies of composite systems are challenging to run for a number of reasons. A task that has high ecological validity for a forensic application (i.e. naming) yields very low response rates, which lead to a floor effect. Most composite systems require extensive training and experience for an operator to become fully skilled in using the system. Experimental control for the skill of an artist is impossible to achieve. Therefore studies using appropriately trained operators are difficult to conduct (but see Frowd *et al.*, 2005a for an example of good practice). An evaluation study comparing EFIT-V with E-FIT using trained operators is underway and will be the subject of a separate report.

Our aim in the experiments reported here was to investigate the applied potential of producing within-witness morphs (i.e. of composites produced by the same witness). Within-witness morphs are predicted to be more effective than individual composites because they will remove random errors, and errors that might be specific to the initial selection of random faces. The effectiveness of within-witness morphs was compared to both between-witness morphs and individual composites using a likeness rating and a naming task. These experiments build on an effect reported by Ness (2005) in an

unpublished PhD thesis. She found an advantage for within-witness morphs of composites created using different systems (e.g., E-FIT, EvoFIT, Sketch and PROfit) compared to individual composites. Between-witness morphs performed better than the within-witness morphs. Our aim was not to carry out an evaluation of EFIT-V in comparison to other systems. However, the data reported here were based on tasks and procedures which have been employed in previous research which facilitates comparison with the existing research literature using other systems.

Comparison was made of composites constructed from memory and with an image in view, to allow the effects of the composite construction system to be separated from limitations of witness memory. In all experiments composites were constructed by people who were not familiar with the target actors. In Experiment 1, participant-judges who were familiar with the actors ranked composites and morphs for their likeness to the target actor. In Experiment 2, the composites and the derived morphs were shown to participants familiar with the target actors for identification by naming the target. The naming rates were conditionalised on participants' ability to name photographs of the target. Experiment 3 removed a potential artifact in the comparison of between- and within-witness morphs using a likeness judgment task. Experiment 4 explored further the means by which morphing yields a benefit, by exploring the effect of morphing on the representation of the internal and external features in composites.

Experiment 1

Method: Creation of EFIT-V composites and morphs

Participants

Twenty (6 male; 14 female) continental European adult student 'witnesses', aged 19 - 44, were paid £15 for a session lasting approximately 2 hours. The recruitment criteria were that they self-reported that they were *unfamiliar* with actors from British TV soap operas from which the target actors in this experiment were selected.

Materials

Ten male target actors (5 Australian and 5 British) were chosen to act as targets. The selected actors were rated as the most recognisable of 20 photographs of soap stars from five different high-rating British TV programmes by 30 British students. All had starred in the programmes for at least six months and their ages were stated to range from 20 – 49. All of the selected actors appeared in *Neighbours* or *Eastenders*. None had beards, glasses or other distinctive features. The names of the selected actors are listed in Annex 1. A high-quality contemporary colour photograph of each of the ten target actors was obtained. A 2-min colour video depicting close-up moving facial views of each actor was recorded and edited using Adobe Pro video editing equipment for playback on an 18" monitor screen.

EFIT-V Software (Vision Metric Ltd., EigenFIT v1.006) was used to create the composites. The software was installed on a computer with a 17" monitor screen. MorphStudio software (The Learning Company) was utilised to merge transformed images.

Procedure

Participant 'witnesses' created four composites of an actor from *view* (with the reference image available throughout), and four composites of a second actor from *memory*. All participants viewed a 2-min colour video of the first target actor. They were

then provided with a brief demonstration of the features available on EFIT-V, which produced images in colour. Condition order was counterbalanced across witnesses so that half created the *memory* set first, half the *view* set first. Finally, each witness viewed simultaneously on the screen all 4 composites she or he had created and ranked them for resemblance to the target from memory.

Participants worked under the guidance of the experimenter who provided advice on use of the software and the features available. The first step with EFIT-V is to select a desired hairstyle from arrays. Specific hair length and colour criteria can be specified. Participants then choose the face that most resembles the target, from a series of arrays of nine photo-realistic synthetic faces each with the selected hairstyle. Participants were encouraged to utilise a number of system functions throughout the process. These features allowed the following procedures: Two or more faces to be blended together. The entire array could be rejected, if none resembled the target more than others, and was replaced with an alternative set of nine faces. Individual internal features (e.g., eyes, nose and mouth), and external features (e.g., face and chin shape) could be 'frozen' (so that the feature did not change with subsequent evolution). Individual features could be moved, or manipulated in width or height. The apparent age of the face could be adjusted. The experimenter's role was to remind participants of the system's features, and to note comments but not to contribute to the process. There was no time limit, although to create all eight composites took approximately 110 min.

MorphStudio software was used to merge images. For within-witness morphs, the four composites produced by each individual witness were merged, so that the composites ranked as most similar contributed the most to the final morph using the

following ratio (40%; 30%; 20%; 10%). For between-witness morphs the four composites by different witnesses self-ranked as 'best' were combined with each contributing the same weight to the final morph (25%; 25%; 25%; 25%). See figure 1. This resulted in 21 images of each actor (16 individual composites – eight from view, eight from memory; four within-witness morphs – two from view, two from memory; and one between-witness morph). The between-witness morph utilised two composites constructed from memory and two composites constructed from view.

Method: Ranking task

Participants

Twenty (5 male; 15 female) adult undergraduate participant-judges aged 19 – 24 were recruited. All self-reported to be regular viewers of *Eastenders* and *Neighbours* and therefore were very likely to be highly familiar with the actors. The participants gained course credit for participation.

Design, materials and procedure

Participant-judges were asked to rank the 21 composites of each of the ten target actors for resemblance to a photograph of that actor. The 21 composites of an actor printed in full colour, sized 80 x 95 mm, were placed randomly on a table alongside a colour reference photograph approximately 50 x 35mm.

The participant placed the photographs in order to reflect their ranking. This took approximately 45 min. The dependent variable was the mean rank of each image type.

Results

Table 1 shows the mean rank out of 21 (collapsed across actors) for each type of composite. 'Actor' was the random factor in all of the analyses reported. The data,

collapsed across experimental condition (constructed from view or memory), were entered into a Kruskall-Wallis test. The type of composite image was significant, χ^2 (2) = 37.37, $p < .001$. Mann-Whitney U tests revealed that the between-witness morphs had a significantly lower mean rank (i.e. a better likeness) than the within-witness morphs ($p < .05$, r = .36), which in turn were ranked significantly lower than the individual composites ($p < .001$, r = .33).

The effect of experimental condition was explored using two further Mann-Whitney tests. It was found that individual composites produced from memory (Mean rank = 11.30) were surprisingly ranked as better than those produced from view (Mean rank = 12.53; $p < .05$, r = 0.16). The difference in the mean ranking of within-witness morphs produced from view or from memory was not significant although the effect size was similar ($p > .2$, r = .17).

A Kruskall-Wallis test was conducted on the mean ranking of the individual composites to investigate whether there was an effect of the order of composite construction (mean rank: first = 12.27, second = 12.50, third = 12.15, fourth = 10.76). The effect of order was not significant, χ^2 (3) = 3.89, $p = .27$. The rank for each actor's individual composite images was analysed as a function of the ranked position given by the witness who constructed the composite, to examine the reliability of witness' assessment of the quality of their own composites. A Kruskall-Wallis test revealed a significant effect, χ^2 (3) = 10.23, $p < .05$. Composites self-ranked as best also tended to be ranked as best by participant-judges (Mean rank = 10.44), followed by those self-ranked second (mean rank = 11.94), third (mean rank = 12.02) and fourth (mean rank = 13.27). Mann-Whitney U tests showed that only the difference between the first and fourth self-

ranked composites were significantly different in the ranks given by the participant-judges ($p < .05$, r = 0.34).

Analysis of the data in Table 1 demonstrated that morphs tended to be rated as better than the individual composites as a group. However, in a forensic situation in which several composites are available, an alternative strategy to creating between- or within-witness morphs would be to select only the 'best' individual composite (i.e. the images self-ranked 1 by each witness). A Kruskall-Wallis test was used to examine whether the morphs were rated as better likenesses by participant-judges to 'best' self-ranked individual composites. This comparison was significant, $\chi^2(2) = 12.98$, $p < .01$. Mann-Whitney U tests revealed that between-witness morphs had a significantly lower mean rank (mean = 5.78) than the within-witness morphs (mean = 8.59, r = 0.36) and 'best' self-ranked individual composites (mean = 10.44; $p < .05$, r = 0.46 for both comparisons). Within-witness morphs were in turn ranked lower than the 'best' individual composites. However, this difference was not significant ($p = .054$, r = 0.22).

In the UK witnesses normally only produce a single composite image. Therefore a further Kruskall-Wallis test was carried out to examine whether participant-judges rated the first composite produced by each witness differently from the morphs. The comparison was significant, $\chi^2(2) = 29.52$, $p < .001$. Mann-Whitney U tests revealed that the between-witness morphs were ranked significantly lower than the within-witness morphs (as reported above). Within-witness morphs were ranked significantly lower than the first individual composite produced **(**mean 8.59 vs. 12.27 respectively $p < .01$, r = 0.47).

Discussion

The data generated by Experiment 1 consisted of rankings of the likeness of images to a photograph of the target person, provided by people who were familiar with the target actors. Between-witness morphs provided better likenesses than within-witness morphs. However, within-witness witness morphs were judged to be better likenesses than the individual composites. Between-witness morphs and within-witness morphs were better likenesses than the self-rated 'best' individual composites and the first composites produced. The effect size of morphing composites within-witnesses was a medium effect; it was compared to the 'best' individual morph (mid-way between the sizes characterised as small and medium) and was a large effect compared to the first-generated composite. Morphing between–witnesses yielded a large effect size.

The composites produced from memory were rated as better likenesses than the composites produced from view, although the effect size was small. This result may appear counter-intuitive, but it is likely to be attributable to the nature of the system of 'evolving' a facial appearance used in EFIT-V. When working from memory the witness had to rely on a recollection and recognition of facial appearance and selected faces that capture some similarity to the target face, although the similarity may be difficult to specify. This is the method which EFIT-V is designed to exploit, capturing the natural face recognition process which exploits configural or holistic properties of faces. In contrast when a photograph is available, witnesses tend to concentrate more on noticeable differences in comparison to specific features. In these circumstances witnesses tend to use the software tools to alter the appearance of features at an earlier stage. The availability of a photograph encourages witnesses to use a feature-by-feature comparison

strategy. Therefore they fail to obtain the full benefit of the holistic processing on which EFIT-V is based.

The statistical analyses were based on data in which the actor was the random factor (i.e. an items analysis), indicating that the advantage for between-witness and within-witness morphs over individual composites generalises reliably across the range of actors who served as targets in Experiment 1. Ranking the likeness of images was an effective experimental method that avoided floor effects and yielded useful data to discriminate between different types of images and experimental conditions. However likeness judgements lack ecological validity for a forensic application. Therefore, the set of images generated for Experiment 1 were shown to a different set of participants for identification by naming in Experiment 2. This experiment has high forensic relevance as it includes composites generated from memory by witnesses who were unfamiliar with the target person, and the ability of people who know the target person to name them was evaluated. This simulates the situation in which publicity is given to facial composites for recognition by police officers or the public.

Experiment 2

Method

Participants

Eight hundred and seventy-six staff, students and visitors to Goldsmiths College were approached to contribute data. However, 226 did not watch *Eastenders* or *Neighbours.* Their data were excluded. Data from 650 participants were retained for analysis.

Materials and Design

The composites used were the same images as described for Experiment 1. Participants familiar with the target actors were invited to attempt to name the images from the composite and morphing procedure. Images were arranged in a series of booklets, each containing ten images, one of each actor. Fifty participants meeting inclusion criteria were assigned to each booklet:

1. Between-witness morphs (BWM);
2. Within-witness morphs memory 1 (produced when completing memory condition first - WWM1)
3. Within-witness morphs memory 2 (produced when completing memory condition second – WWM2)
4. Within-witness morphs view 1 (produced when completing view condition first – WWV1)
5. Within-witness morphs view 2 (produced when completing view condition second – WWV2).

Eight further booklets were constructed containing individual composites produced from *memory* only by participants. The booklets were constructed to contain composites from one of the eight following conditions:

M11 – M14: First to fourth individual composites created when taking part in memory condition first.

M21 –M24: First to fourth individual composites created when taking part in memory condition second.

The individual composites produced from view were not included in this experiment. The dependent variable was the Conditional Naming Rate (CNR), calculated

by dividing the number of correct identifications made to each image by the number of participants who could subsequently name the target actor from a photograph.

Procedure

Individual participants were asked to examine one of the booklets to see if they could identify any of the 'celebrities'. Next, they were told that the images were of Australian or British soap actors. If participants claimed not to view this type of programme, or once the procedure was complete they claimed not to specifically view *Eastenders* or *Neighbours*, their participation was terminated. The remainder were asked to inspect the images again and a record was kept of the number they correctly named, or for whom they managed to produce an individuating semantic description. Real photographs of the target actors were then shown and the participants were asked to identify the actors from their photograph.

Results

Of the 650 participants included in the final analyses, 62 regularly watched *Neighbours* only, 192 watched *Eastenders* only. The remainder watched both. Participants who claimed to watch a programme were always 100% accurate in naming photographs of actors from that programme. Only 5 participants (0.8%) spontaneously produced the correct name of a celebrity, prior to being advised that the pictures were of soap actors. Figure 2 shows the distribution of participants as a function of the proportion of actors correctly identified after receiving the soap opera cue. All participants who viewed between-witness morphs named at least one actor, one participant correctly named nine. In contrast, 21.5% of participants (n = 86) who viewed individual

composites and 12.0% (n = 24) who viewed within-witness morphs were unable to name any of the images.

A one-way independent-measures ANOVA of the CNR data with image type as a factor was significant, $F(2, 647) = 58.55$, $p < .001$, $\eta^2 = .15$. Games-Howell post-hoc tests showed that the between-witness morphs (M = 43.8%; SD = 17.8) were identified significantly more often than within-witness morphs (M = 32.1%; SD = 20.5), which were also identified more often than individual composites (M = 20.3%; SD = 15.9; $p <$ .001 for all comparisons).

A second analysis compared naming of within-witness morphs of composites constructed with the target actor's photograph in view to within-witness morphs of composites constructed from memory. The effect of the order of these experimental conditions was also examined. A 2 (view vs. memory) x 2 (first vs. second) between-participants ANOVA on the CNR found that the main effects of condition, $F(1, 196) = 1.04$, $p = .31$, $\eta^2 = .05$ and order, $F(1, 196) = .28$, $p = .60$, $\eta^2 = .01$, were not significant. However, there was a significant interaction between condition and order, $F(1, 196) = 8.91$, $p < .005$, $\eta^2 = .043$. Simple main effects found that more within-witness morphs were identified from composites made in the first set when they was made from view than when they were made from memory (35.6% vs. 30.0%), although this difference was not significant, $F(1, 196) = 1.93$, $p = .17$, $\eta^2 = .01$. In contrast, significantly more within-witness morphs were recognised from participants' second set composites when they were constructed from memory than from view 37.0% vs. 25.6%, $F(1, 196) = 8.01$, $p <$ .005, $\eta^2 = .04$.

A 2 (condition order: from memory first vs. second) x 4 (composite construction order: first, second, third, fourth) between-participants ANOVA was conducted on the CNR to individual composites to examine whether there were any effects of the order of production within an experimental condition. The main effects of condition order, $F(1, 392) = 0.85$, $p = .36$, $\eta^2 = .002$; composite construction order, $F(3, 392) = 2.20$, $p = .09$, $\eta^2 = .017$ and the interaction, $F(3, 392) = 2.52$, $p = .058$, $\eta^2 = .019$ were not significant.

To examine the reliability of witnesses to assess the quality of their own composites, the CNR for each actor's individual composite images was analysed as a function of the self-ranked position. Nine of the images self-ranked as their best creation by the 20 witnesses, seven of the second best ranked, three of the third best, and one image ranked worst were associated with the highest CNR for an individual actor. A repeated-measures one-way ANOVA on these data was significant, $F(3, 57) = 4.95$, $p < .05$, $\eta^2 = .21$. Bonferonni-corrected paired comparisons found that there were no differences in the CNR between composites self-ranked first (CNR = 23.0%, *SD* = 15.3), second (21.9%, *SD* = 15.0) and third (20.8%, *SD* = 12.6; $p > .1$). However, the CNR of those self-ranked fourth by witnesses was significantly lower than all others (15.0%, *SD* = 11.2; $p < .05$ for all comparisons).

A further test compared the CNR of the morphs to the CNR of the 'best' witness self-ranked individual composites (mean = 18.6%, SD = 29.9**)**. To ensure parity of conditions, only participants who viewed *both* TV soaps were included in this analysis[2]**.** An independent-measures one-way ANOVA conducted on these data was significant, $F(2, 393) = 13.84$, $p < .001$, $\eta^2 = .07$. Games-Howell post-hoc tests found that although between-witness morphs were identified more often than within-witness morphs, the

difference was not significant ($p > .2$). However, between- and within-witness morphs were both identified significantly more often than the self-rated 'best' witness individual composites ($p < .01$).

A final analysis compared CNR of the morphs with the 'first' individual composites only (mean = 17.5%, SD = 13.3). A one-way independent measures ANOVA on these data was significant, $F(2, 347) = 38.54$, $p < .001$, $\eta^2 = .18$. Games-Howell post hoc tests revealed that between-witness morphs were identified significantly more often than within-witness morphs which in turn were identified significantly more often than 'first' individual composites ($p < .01$ for all comparisons).

The statistical analysis of Experiment 2 took the participant-judges as the random factor. The results show that the advantage for between-witness and within-witness morphs over individual composites generalise reliably across the participants, but the data are collapsed across actors. Therefore a one-way repeated measures ANOVA taking actor as the random factor was carried out to compare the CNR for between-witness morphs, within-witness morphs and individual composites. The assumption of sphericity was met and the effect of image type was significant, $F(2,18) = 9.68$, $p=.001$, $\eta^2 = .52$. Bonferroni-corrected comparisons showed that a greater proportion of between-witness morphs and were named than were individual composites, $p =.004$, Cohen's d = 1.22. The difference between individual composites and within witness morphs was not significant using the Bonferroni-corrected comparison, $p = .10$, but the effect size was large, Cohen's d = 0.80. The difference in the proportion of between-witness and within-witness morphs named was not significant, p = .16, but the effect size was medium, Cohen's d = .70.

Discussion

Experiment 2 found an advantage for naming between-witness morphs compared to both within-witness morphs and individual composites. Within-witness morphs were better recognised than the individual composites. The data demonstrate a medium to large effect size of morphing composites constructed by witnesses who did not know the target, in a task that required participants to name or otherwise uniquely identify composites of targets who were known to them. Morphing composites has been demonstrated to be forensically useful even when the only composites available are all produced by a single witness. Both between-witness and within-witness morphs were better recognised than either the composite rated at the most similar to the target by the witness or the first composite that the witness constructed.

Witnesses who constructed composites 'from view' first, and 'from memory' second, produced more recognisable within-witness morphed composites. Although this interaction of condition (view vs. memory) and order had a statistically significant effect on naming within-witness morphs, the effect size was small. It is unclear whether this was an effect of the order of different conditions or whether these particular witnesses were simply more competent at the task.

Experiment 2 provided the first empirical evaluation of people's ability to name EFIT-V composites. Caution must be exercised when comparing naming rates across studies, because the individual circumstances of each experiment is unique. The most forensically-relevant comparison is that of naming rates across studies in which the witness was unfamiliar with the target, but the person viewing the composite was familiar with the target. Using Pro-Fit (a commercial competitor of E-Fit in the UK), Bruce *et al*. (2002) obtained recognition rates of 16 – 22% of individual facial composites of staff in a

university department in a context that provided a strong contextual cue to naming. Frowd *et al*. (2005) found a naming rate of approximately 3% for composites of celebrities constructed using Evo-Fit and FACES. However in Frowd *et al*.'s experiment there was no contextual cue to aid naming and there was a delay of 2 days between the witness viewing the (unfamiliar) target person and constructing the composite. In Experiment 2 the individual composites were identified by 20% of participants given the cue "an Australian or British soap actor". This naming rate is very similar to that reported by Bruce *et al*. (2002). The between-witness morphs in Experiment 2 were identified by 44% of participants in Experiment 2 compared to 28% in Bruce *et al*.'s study. A conservative conclusion is that EFIT-V produces composites that are at least as well recognised as the best performance observed for the existing systems.

There are two factors that might contribute to the advantage of between-witness morphs over within-witness morphs in Experiment 2 (44% vs. 32% named). First, errors between witnesses are more likely to be uncorrelated than errors made by the same witness generating a composite of the same target multiple times. If a witness incorrectly recalls a feature, he or she is likely to attempt to select the erroneous feature in each composite produced. Therefore, errors are more likely to be reduced by the morphing process involved when generating between-witness morphs. Second, between-witness morphs were selected from the composite self-rated by the witness as the best likeness of the four composites produced. In contrast, within-witness morphs are weighted morphs of a best, a second-, a third- and a fourth-ranked composite. Therefore, the advantage for between-witness morphs could be attributed to either the diversity of the images produced by different witnesses or to the quality of the composites from which the morph

was created. The aim of Experiment 3 was to investigate whether there was an advantage for between-witness morphs over within-witness morphs when the latter factor is eliminated.

Another difficulty for the comparison of between-witness and within-witness morphs in Experiment 1 and 2 is that the between-witness morphs were morphs of two composites constructed from memory and two composites constructed with the target in view. In comparison the within-witness morphs were morphs of 4 composites constructed either from memory or from view. This factor seems unlikely to have strongly influenced the results. Composites constructed in view were rated as poorer likenesses than were composites constructed from memory. Therefore any confounding effect would have reduced the differences between the two types of morphs by tending to render between-witness morphs poorer likenesses. In Experiment 3 this confound was removed. All composites were constructed from memory.

Experiment 3

Method

Design

In each trial of Experiment 3 participant-judges viewed a facial composite and a photograph of the target actor simultaneously on a computer monitor and provided a similarity rating. A 3 (image type: individual composites, within-witness morphs, between-witness morphs) x 5 (actor) repeated measures design was used. The dependent variable was a similarity rating (using a scale of 1-10).

Participants

There were twenty participant-witnesses (5 male; 15 female; mean age = 29.0 years) who constructed a new set of facial composites for use in Experiment 3. None were familiar with the television soap *Hollyoaks* (Channel 4) or any of the target actors. There were thirty participant-judges aged 18 – 25 years (4 male; 26 female) who were all regular viewers of *Hollyoaks* and therefore familiar with the target actors.

Materials construction

Television still color close-up facial images (eight per actor) were obtained for five actors (Annex 1) from the television soap *Hollyoaks*. The actors were white Caucasian males aged 20-30 with short hair and no beards, glasses or hats. The eight shots of each actor with a neutral expression included one full-face and two profile (left and right) poses. The remaining images were three-quarter views. The images of each target actor were presented on two A4 (210 mm x 297 mm),sheets.

After a demonstration of EFIT-V using a female face, each participant witness was randomly assigned to one of the five unfamiliar targets. After viewing the set of eight photos of that target for two minutes they created four consecutive composites from memory. The whole procedure took approximately two hours. Each participant ranked the composites they had created from best to worst at the end of the production process. For each of the five target actors, four different participant-witnesses created four images each. This resulted in 16 individual composite images of each of the targets. Morph Studio software was then used to morph composites as described in Experiment 1.

For within-witness morphs, the method of construction was the same as in Experiments 1 and 2, in that the composites ranked as most similar contributed the most to the final morph using the following ratio (40%; 30%; 20%; 10%). This resulted in four

within-witness morphs of each actor. However, the method used to create between-witness morphs differed from that used in Experiments 1 and 2. The highest self-ranked composite contributed by one witness contributed 40% to the between-witness morph, the second highest composite by a different witness contributed 30%, the composite self-ranked third by a third witness contributed 20%, with the composite self-ranked as worst by a fourth witness contributed 10%. This method was rotated across the four witnesses to produce four between-witness morphs of each actor, so that each witness contributed equally across the four morphs.

This method resulted in 24 images of each target actor (120 in total). Sixteen of these were individual composites created by the four witnesses allocated to each target (4 per witness). Four were within-witness morphs (created by combining 4 weighted composites of each witness) and four were the between-witness morphs (created by combining 4 weighted composites produced by the 4 different witnesses).

Procedure

Participant-judges were presented with the 120 final images on a computer monitor using PowerPoint with one composite image per slide on the right-hand side of the screen. Two of the best still shots (from the eight viewed by the participant-witnesses in stage one) of each target always appeared on the left-hand side of the screen to serve as a reminder of their appearance. Participant-judges rated each composite for similarity to the target actor on a scale of 1-10. The 24 images for each actor were randomly ordered and the order in which the images of each actor were presented was counterbalanced.

Results

The mean similarity ratings for each image type of the five actors were calculated and the results are presented in Figure 3. A 3 (image type: individual composites, within-witness morphs, between-witness morphs) x 5 (Actors 1 - 5) repeated measures ANOVA was conducted on the similarity ratings. The main effect of image type, Mauchly's W = .52, χ^2 (2) = 18.43, p < .001, ε = .68, and the interaction, Mauchly's W = .02, χ^2 (35) = 102.31, p < .001, ε = .58, violated the assumption of sphericity so a Greenhouse-Geisser adjustment was used. The main effect of image type was significant, $F(1.35, 39.13) = 71.16$, $p < .001$, $\eta^2 = .71$; Ratings were higher for between-witness morphs ($M = 3.81$, $SD = 1.23$) than for within-witness morphs ($M = 3.06$, $SD = .90$) which were also higher than for individual composites ($M = 2.38$, $SD = .67$, p < .001 for all comparisons). The main effect of actor was also significant, $F(4, 116) = 24.62$, $p < .001$, $\eta^2 = .46$; with mean ratings varying across the actors. However, these effects were mediated by a significant interaction, $F(4.66, 135.14) = 3.63$, $p = .005$, $\eta^2 = .11$. Bonferroni-corrected paired comparisons on the effects of image type within each the five actors revealed that for three of the actors (Actors 1, 2 and 3) the differences were consistent with those reported for the main effect of image type, in that between-witness morphs were rated as better than within-witness morphs which were also rated as better than the mean individual composites ($p < .01$ all comparisons). For the other two actors (Actors 4 and 5) the morphs were rated as significantly better than the individual composites ($p < .01$). However, the difference between between-witness morphs and within-witness morphs was not significant ($p > .05$).

To examine whether the morphs were also rated as better than the best self-ranked composites a similar 3 (image type) x 5 (actor) ANOVA was conducted in which the

overall mean composite rating was replaced with the best self-ranked image (M = 2.57, SD = 0.71). The main effect of image type, Mauchly's W = .48, χ^2 (2) = 20.76, p < .001, ε = .66, and the interaction, Mauchly's W = .06, χ^2 (35) = 73.04, p < .001, ε = .64 violated the assumption of sphericity so a Greenhouse-Geisser adjustment was used. The main effect of image type was significant, $F(1.31, 38.07) = 48.34$, $p < .001$, $\eta^2 = .62$; Ratings were higher for between-witness morphs than for within-witness morphs which were also higher than for individual composites ($p < .001$ for all comparisons). The main effect of actor was also significant, $F(4, 116) = 23.83$, $p < .001$, $\eta^2 = .45$; with mean ratings varying across the actors. However, these effects were mediated by a significant interaction, $F(2.39, 147.82) = 3.63$, $p < .05$, $\eta^2 = .09$.

Bonferroni-corrected post-hoc analyses on the effects of image type within each of the five actors revealed that for two of the actors (Actors 3 and 5) the differences were again consistent with those reported for the main effect of image type, in that between-witness morphs were rated as better than within-witness morphs which were also rated as better than the mean individual composites ($p < .01$ all comparisons). For two further actors (Actors 1 and 4) the between-witness morphs were rated as significantly better than the within-witness morphs and the individual composites ($p < .01$). However, the difference between within-witness morphs and the individual composites was not significant ($p > .05$) .With the final actor (Actor 5) the difference between the between-witness morphs and individual composites were significant ($p < .01$). However, all other differences were non-significant ($p > .05$).

Discussion

Experiment 3 showed that between-witness morphs produced better likenesses than within-witness morphs when both types of morphs were constructed from

composites selected and weighted in exactly the same manner. Experiment 3 eliminated a possible interpretation of Experiments 1 and 2 that between-witness morphs only produced better likenesses than within-witness morphs because between-witness morphs were morphs of the best composite from each witness rather than a weighted morph of the first to last rated morphs. The morphs used in Experiment 3 all consisted of 40% of a composite rated the best likeness by a witness, 30% of a composite rated as the second-best likeness by a witness, 20% of a third-rated likeness and 10% of a worst-rated likeness. The results of Experiment 3 show that morphing across witnesses produced better likenesses. This effect is consistent with an explanation that errors across witnesses are more likely to be unrelated than are errors in composites produced by the same witness. Therefore morphing across witnesses is more effective in averaging out uncorrelated errors. Experiment 3 also replicated the superior likenesses of within-witness morphs compared to individual composites. Within-witness morphs are at least as good a likeness as the best individual composite. Therefore the utility of within-witness morphs for police investigations has been demonstrated again. Experiment 3 also eliminated a confound in Experiments 1 and 2 in which only the between-witness morph were constructed from two composites that were produced with the target face in view and two composites produced from memory.

Analysis of Experiment 3 showed a large effect size of morphing, and a large effect of actor. The advantage of between-witness and within-witness morphs over individual composites was broadly consistent across all five actors, but likeness ratings differed substantially across actors as would be expected. The interaction of morphing with actor produced a moderate effect size. Not all of the ‘main effect’ comparisons were

reliable for all actors individually. However, minor differences for individual actors are to be expected, and no actor showed a substantial deviation from the mean effects.

The utility of a facial composite relies on its effectiveness in prompting recognition by people who know the target person depicted (for example by circulation to police officers who may recognise a previous offender or by a viewer of TV news or a crime programme). Cognitive processes involved in recognising familiar faces differ in some important respects from those involved in recognising faces only seen once previously (unfamiliar faces). Whilst recognition of unfamiliar faces tends to be dominated by the external features (e.g. hairstyle and texture, face shape), the internal features of a face (eyes, nose, mouth) are relatively more important in recognising familiar faces (Ellis, Shepherd and Davies, 1979; Young, Hay, McWeeny, Flude & Ellis, 1985). Frowd, Bruce, McIntyre and Hancock (2007) reported evidence that the difficulty of identifying facial composites of familiar persons may be attributed to an inability of the facial composite systems to reproduce the internal features effectively. The external features of facial composites of celebrities were better matched to a target face than were the internal features even when the composites were constructed by witnesses who were familiar with the target celebrity. Therefore, Frowd *et al.* argued that the poor recognition of internal features was attributable to the process of composite construction itself.

The aim of Experiment 4 was to investigate whether morphing composites, both between-witnesses and within-witnesses, produced better quality images because they improved the representation of the internal features. As argued previously, if the construction process introduced random or idiosyncratic errors, morphing across images that contain different errors in the internal features would produce images that are more

similar to the target face by 'averaging out' errors but maintaining consistent feature information. It was hypothesised that the external features would benefit less than the internal features from morphing because there are limited options for selecting a hairstyle. A set of four composites of a target person may have the same hair because the witness or witnesses made the same selection. Similarly, morphing within-witnesses is likely to benefit the quality of the image less than morphing between-witnesses because idiosyncratic errors are more likely to be correlated across images in the within-witness morph. In Experiment 4 the between-witness morphs were produced from the best-ranked image produced by four witnesses. This method was selected because it is the method that would be most appropriate to use in a criminal investigation because it would yield the best possible quality between-witness morph. Therefore this method is of most applied relevance.

Experiment 4

Method

Design

In each trial of Experiment 4 participant-judges viewed a facial composite and a photograph of the target actor simultaneously on a computer monitor and provided a similarity rating. A 3 (image type: individual composites, within-witness morphs, between-witness morphs) x 3 (feature presentation: internal features, external features, whole faces) repeated measures design was used. The dependent variable was the similarity rating (using a scale of 1-10).

Participants

There were forty participant-judges (14 male; 26 female; Mean age = 25.68 years). All were regular viewers of *Hollyoaks* and therefore familiar with the target actors.

Material construction.

The individual composites all produced from memory as described in Experiment 3 were used in Experiment 4. Composites were morphed as described in Experiment 1, to produce 21 images of each target actor (105 in total). Sixteen of these were individual composites created by the four witnesses allocated to each target (4 per witness). Four were within-witness morphs (created by combining four weighted composites of each witness) and one image was the between-witness morph, created using the 4 top-ranked individual composites of each witness (as in Experiments 1 and 2).

The whole face images used in Experiment 4 were the 105 images described above. To create the internal feature images, an oval mask was inserted over the whole images obscuring the external features, so that only the internal features of the face could be viewed. The oval mask did not include any hair for any of the actors. The same oval mask and was used to produce the external feature images. However, this time the internal component of the oval was obscured, and the external components were visible (Figure 4). This process resulted in a total of 315 images.

Procedure

Participant-judges rated each of the 315 images for similarity to the target actor on a scale of 1-10 using the method described for Experiment 3. The images were presented in three blocks of 105 images of each feature presentation (i.e. whole faces, internal features or external features). The order of presentation of the blocks was counter-

balanced. Within each block all 21 images of one actor was shown before all 21 images of the next actor, and so on. The 21 images for each actor were presented in a different random order for each participant. The order in which actors were presented was counterbalanced.

Results

The mean similarity ratings for each image type of each actor were calculated. First, an analysis of whole face images only as function of image type was carried out to establish whether the results were consistent with those of the previous experiments. A repeated-measures ANOVA was conducted on the mean ratings for the whole images only. The data violated the assumption of sphericity, Mauchly's W = .69, χ^2 (2) = 14.13, p < .001, ε = .76 so a Greenhouse-Geisser adjustment was used. The effect of image type was significant, $F(1.53, 59.52) = 150.14$, $p < .001$, $\eta^2 = .79$. Planned comparisons found that between-witness morphs (mean = 5.67) were rated more similar to the target photographs than within-witness morphs (mean = 4.49), which were rated more similar than the 'best' self-ranked individual composites (mean = 3.90; p < .001 for both comparisons).

Figure 5 depicts the mean ratings as a function of image type and feature presentation, collapsed across the five actors. A 3 (image type: individual composites, within-witness morphs, between-witness morphs) x 3 (feature presentation: internal features, external features, whole faces) x 5 (actor: 1 - 5) repeated measures ANOVA was conducted on the similarity ratings. As appropriate the Greenhouse-Geisser adjustment is reported for all main effects and interactions, for which the assumption of sphericity was violated.

The main effect of image type, Mauchly's W = .63, χ^2 (2) = 17.39, p < .001, ε = .73, was significant, $F(1.46, 57.05) = 189.27$, $p < .001$, $\eta^2 = .830$; Ratings were higher for between-witness morphs ($M = 5.29$, $SD = 1.03$) than for within-witness morphs ($M = 4.50$, $SD = .95$) which were also higher than for individual composites ($M = 3.76$, $SD = .96$, $\underline{p} < .05$ for all post-hoc comparisons). The main effect of feature presentation was also significant, $F(2, 78) = 7.76$, $p < .001$, $\eta^2 = .17$; the rating for internal images ($M = 4.30$, $SD = .99$) were lower than those for both external images ($M = 4.67$, $SD = .92$) and whole images ($M = 4.58$, $SD = 1.10$; $\underline{p} < .05$ for both post-hoc comparisons). The latter two conditions did not differ ($p > .2$). The main effect of actor, Mauchly's W = .47, χ^2 (9) = 28.13, p < .001, ε = .73, was significant, $F(2.85, 111.24) = 30.24$, $p < .001$, $\eta^2 = .44$; an expected effect of the difference in memorability of the actor's faces (actor 1; M = 3.36, SD = 1.18; actor 2; M = 5.13, SD = 1.38; actor 3; M = 4.28, SD = 1.09; actor 4; M = 3.97, SD = 1.05; actor 5; M = 4.74, SD = 1.20).

The experimentally important two-way interaction between image type and feature presentation, Mauchly's W = .28, χ^2 (9) = 47.59, p < .001, ε = .66, was significant, $F(2.65, 103.45) = 29.47$, $p < .001$, $\eta^2 = .43$. Bonferroni-corrected post-hoc analyses on the simple effects of feature presentation within the three image types revealed that for the individual composites, external features received significantly higher ratings than both internal and whole images ($p < .01$) which did not differ ($p > .2$). For the within-witness morphs, external images received higher ratings than internal images ($p < .01$); no other differences were significant ($p > .2$). For the between-witness morphs, higher ratings were given to the whole images than to the external and internal images ($p < .01$), which did not differ ($p > .2$).

The interaction between image type and feature presentation was explored further by splitting the data into smaller ANOVAs. First, two separate 3 (image type: between-witness morphs, within-witness morphs, individual composites) x 2 (feature presentation) repeated-measures ANOVAs were carried out. Data from only the whole face and internal feature conditions were entered into one ANOVA, and from only the internal and external feature conditions were entered into a second ANOVA. The Greenhouse-Geisser correction for violation of the assumption of sphericity was applied to both analyses, Mauchly's W = .57, χ^2 (2) = 21.72, p < .001, ε = .70 and Mauchly's W = .55, χ^2 (2) = 22.72, p < .001, ε = .69 respectively. The effect of image type was greater for whole faces than for internal features, $F(1.39, 54.34) = 14.86$, $p < .001$, $\eta^2 = .28$, and was greater for internal features than for external features $F(1.38, 53.8) = 19.14$, $p < .001$, $\eta^2 = .33$. Thus morphing improved the similarity ratings of whole faces more than of internal features, which in turn were improved more than the ratings of external features. The likeness ratings of the internal and external features of morphed and individual composites were compared separately for within-witness morphs and between-witness morphs in two further 2 x 2 ANOVAs. Each ANOVA had two levels of feature presentation (internal vs. external features) and two levels of image type. In the first ANOVA the levels of image type were individual composites vs. within-witness morphs, and in the second ANOVA the levels were individual composites vs. between-witness morphs. In both ANOVAs the assumption of sphericity was met. There was a significant interaction, $F(1, 39) = 31.32$, $p < .001$, $\eta^2 = .44$ for the comparison with within-witness morphs and $F(1, 39) = 28.53$, $p < .001$, $\eta^2 = .42$ for the comparison with between-witness

morphs. Thus, both within- and between-witness morphing improved the likeness of the internal features more than it improved the likeness of external features.

Returning to the results of the global ANOVA, the two-way interaction between image type and actor, Mauchly's W = .02, χ^2 (35) = 152.25, p < .001, ε = .55, was significant, $F(4.36, 170.14) = 5.63$, $p < .001$, $\eta^2 = .13$; post hoc analyses found variations in the relative ratings given across actors within each of the three image types. These data are shown separately for each actor in Table 2. For all actors, the between-witness morphs were rated as best, the individual composites as worst, consistent with the main effect of image type.

The two-way interaction between feature presentation and actor, Mauchly's W = .14, χ^2 (35) =71.72, p < .001, ε = .66, was also significant, $F(5.32, 207.33) = 6.69$, $p < .001$, $\eta^2 = .15$. Post hoc analyses found that for four of the five actors the pattern was consistent with the main effect of feature presentation, in that internal images were rated as worst ($p < .05$). However, for one actor (Actor 2) the external images were rated as worst ($p < .05$).

The two-way interactions with the effect of actor were mediated by a significant three-way interaction, Mauchly's W < .001, χ^2 (135) = 370.80, p < .001, ε = .49; $F(7.83, 305.39) = 3.95$, $p < .001$, $\eta^2 = .09$. Post-hoc analyses found that for the individual composites and within-witness morphs the influence of the five individual actors was consistent with the two-way interaction between image type and feature presentation (above). However, there were variations for two of the five actors with the between-witness morphs, in that ratings for whole faces and external features for one actor did not significantly differ, and the internal features were rated as significantly worse. In contrast,

ratings for whole faces and internal features for a second actor did not significantly differ, and the external features were rated as significantly worse. For the remaining three actors, higher ratings were given to the whole between-witness morph images than to the external and internal images as consistent with the two-way interaction

Discussion

Using a similarity rating to a photograph on a 10 point scale, Experiment 4 found an advantage for morphs over individual composite with a large effect size. This is consistent with results of Experiment 1 using a likeness ranking task. As in Experiment 1, between-witness morphs were also more similar to the target than the within-witness morphs.

Likeness ratings for the individual composites showed that the external features alone were rated as more similar to a photograph of the target than were the internal features alone. Interestingly the external features were rated more similar to the target than the whole face. Whole face presentation would enable holistic or configural properties to be processed, which are known to be important in face recognition (e.g. Leder and Bruce, 2000; Tanaka & Farah, 1993). When the external features were combined with erroneous internal features, configural processing may allow a different identity to be perceived (Young, Hellawell & Hay, 1987), making whole faces appear more dissimilar to the target than the external features alone. The result suggests that, notwithstanding the familiarity of the participant judges with the target person, similarity to the target was dominated by the external features. This result replicates an effect reported by Frowd *et al.* (2007) using a different composite system. The dominance of external features can be explained by inaccuracy in reproducing the internal features in

the composites. Lack of familiarity of the witness-participants who produced the composites with the target person, and limitations of the composite construction process, may both be factors that may contribute to the poor construction of a likeness of the internal features. The observation that the construction process is dominated by external features is consistent with previous research that showed that the external features are the most salient in recognition of unfamiliar faces (e.g., Ellis *et al*., 1979).

Morphing four composites produced by the same witness increased the similarity of all feature presentations (whole, internal and external) to the target. The external features of within-witness morphs were more similar to the target than the internal features alone, but for within–witness morphs the similarity of the whole face and the external features did not differ significantly. Thus morphing improved the similarity to the target of both internal and external features, by averaging out random variation or error that differed between subsequent attempts to produce a composite. The substantial improvement in the similarity-to-target rating of the internal features of within-witness morphs (from 3.49 for individual composites to 4.33 for within-witness morphs) removed the effect of poor internal features impairing the likeness of whole faces compared to the likeness of external features only that was observed for the individual composites.

Morphing composites produced by different participants increased similarity to the target more than within-witness morphing, as found in Experiments 1 and 3. There was no difference in the similarity to the target of the internal and external features of between-witness morphs, but the whole face of between-witness morphs was most similar to the target. The good likeness of both the internal and external features of

between-witness morphs combined to produce an even stronger resemblance for whole faces. This effect could be attributed to the emergence of resemblance from the configural or holistic properties of the whole face which are important to human face recognition.

A comparison between individual composites and between-witness morphs showed that the increase in similarity to the target due to morphing was greater for internal features than for external features. The difference gave a large effect size. A similar comparison between individual composites and within-witness morphs also showed that the morphing process created a greater increase in similarity to the target for internal features than for external features. The size of this effect was large. External features probably benefit less from morphing for two reasons. First, the witnesses are likely to have a relatively good memory of the external features of the unfamiliar faces they tried to reconstruct. Second, due to the wide variety of hair styles, texture and color there are likely to be a limited number of plausible alternative hairstyles available for selection in the facial composite system, therefore giving less variability in the individual composites that contribute to the morph.

It is unsurprising that there is a large effect size across different actors in the similarity ratings of the different feature presentations and the effect of morphing. Some faces will be easier to recognize than others. For some the external features may be more distinctive, for other faces the internal features may be relatively more distinctive. Therefore variation is to be expected. Whilst the effect size of morphing and of feature presentation were large, the size of effects that interacted with actor was medium. The

analysis of the effects across different actors showed that the broad trends of the data were consistent across faces.

In conclusion, morphing benefits the representation of the internal features more than it benefits the representation of the external features. Nevertheless, both internal and external features are better represented in morphs than in the original composites. Whole faces benefit more from morphing than do the internal features alone.

General Discussion

The new generation of facial composite systems that use a genetic algorithm to evolve a photographic-quality artificial facial appearance have introduced an approach that is easier for the witness to use, and allows more cognitively relevant manipulation of images (e.g., ageing, caricature manipulations of masculinity or femininity). The experiments reported here have demonstrated a clear advantage of morphing composites produced by different witnesses or by a single witness. Morphs of composites produced by four different witnesses were named by over 40% of participants, and morphs of composites produced by a single witness were named by over 30% of participants, compared to just 20% who named the individual composites. The experiments reported in this paper provided the first demonstration of a statistically reliable advantage in a naming task for between-witness morphs of composites made by witnesses who were unfamiliar with the target person. The experiments also provide the first report of statistically reliable advantage for within-witness morphs on likeness judgments and a naming task using a single composite system. Furthermore the experiments used EFIT-V - a system that is now widely used by police in the UK – and demonstrated it to be effective in producing naming rates that are comparable or better than naming rates

observed from other systems in broadly similar conditions. These data suggest that the technology behind EFIT-V, a facial composite system that evolves artificial facial appearance under the control of a genetic algorithm, has reached a level of sophistication that produces realistic, recognizable representations of familiar faces. There is no reason to believe that the benefits of morphing reported here are specific to EFIT-V or the new generation of systems that use a genetic algorithm to search a facial appearance space. The benefit of between-witness morphs has been observed using a conventional software system (Bruce et al., 2002) and of within-witness morphs of composites produced by different systems (Ness, 2005).

The comparison of between-witness and within-witness morphs in Experiments 1 and 2 were limited by two factors. The advantage for between-witness morphs over within-witness morphs could be due to an advantage of morphing images from different witnesses because errors across witnesses were less likely to be correlated. However, the advantage might arise because the between-witness morphs were derived from four best-ranked images. A second factor is that between-witness morphs, but not within-witness morphs were made from two composites constructed from memory and two composites with the target face in view. The 'in view' composites were less good likenesses than the 'from memory' composites. Therefore, any effect of this confound would have worked against the advantage for between-witness morphs. Neither of these factors affected the comparison of within-witness morphs with individual morphs, which was the main novel comparison in these experiments. Both of these problems were removed in the design of Experiment 3. All composites were produced from memory and all morphs were produced from self-ranked composites combined in the same proportions. These data

clearly show a large effect of morphing composites from different witnesses enhances the likeness of the images most effectively. Morphs of images produced by the same witness produce better likenesses than the individual composites. In conclusion, morphing appears to work by averaging out errors in the composites and works best when the errors are uncorrelated.

Another method of reducing error is to make an anti-caricature of a face. This process will make all facial features less distinctive, whether the features are correct or incorrect. Frowd, Bruce, Ross, McIntyre and Hancock (2007) found that a slight anti-caricature of a composite produced the best likeness of a target face, consistent with the role of reducing distinctive errors. The averaging effect of both morphs and anti-caricatures appear to enhance the likeness of composites produced by witnesses (Hasel & Wells, 2007).

Frowd, Bruce and Hancock (2008) reported that the simple expediency of blurring the hair during the selection of the remaining features when constructing a composite improves the quality of facial composites. The benefit is believed to occur because blurring the external features focuses attention on the internal features. The effect of morphing on the representation of internal and external features was explored in Experiment 4. The likeness of both internal and external features was improved by morphing. External features were better represented than internal features in the individual composites. The likeness of whole-face individual composites appeared to be impaired by holistic or configural processing of inaccurate internal features. Between-witness morphs improved the likeness of internal and external features more than within-witness morphs, which in turn produced better likenesses than the individual composites.

Thus morphing, especially across different witnesses, reduced error in the representation of both internal and external features.

Research over many years has revealed generally rather disappointing results on the utility of facial composites. In addition to the advantages of between-witness and within-witness morphs explored here, a number of techniques have been demonstrated recently to improve the recognition rate of facial composites substantially. Simply showing a number of composites constructed by several different witnesses (Brace, Pike & Kemp, 2000, Bruce *et al.*, 2002) or the same witness (Frowd, Bruce, Plenderleith & Hancock, 2006) increased recognition rates compared to showing a single composite image. A slight anti-caricature improves likeness, as does showing a range of caricatures (Frowd *et al.*, 2007) The ability of composite systems to accurately reproduce hairstyle is limited. A new style of cognitive interview that encourages holistic processing of faces has recently been demonstrated to lead to production of better likenesses (Frowd, Bruce, Smith & Hancock, 2008). Many of these results have been demonstrated using forensically-relevant experimental designs. The future for the forensic application of facial composite system looks much more encouraging than it did just a few years ago. The challenge for future research will be to evaluate these techniques under operational conditions and to discover the optimum combination of such techniques for operational use.

Implications for policy

The experiments reported here suggest that if a facial composite is produced during a criminal investigation to help identify a suspect, there is an advantage for the witness to produce multiple composites which can be morphed. If multiple witnesses are

available a between-witness morph would be most effective, but a within-witness morph would be more effective than a single composite when there is only a single witness and it is not possible to make a between-witness morph. Importantly the data reported here showed that a within-witness morph will be as a good a likeness as the best or the first–produced composite. Therefore, using a within-witness morph does not run the risk of creating a sub-optimal likeness. Within-witness morphs may be more acceptable for legal reasons than between-witness morphs. A within-witness morph is the product of a single witness. Morphing between witnesses, who have had different opportunities to view the perpetrator, produces an image that may differ substantially from any image produced by any single witness. The witnesses may even have seen different people, one of which is mistakenly believed to be the perpetrator. This objection cannot be made of a within-witness morph. Although the data clearly show that a between-witness morph is the most effective image, within-witness morphs can be used in more cases and are likely to be less controversial from a legal point of view. The Association of Chief Police Officers (ACPO) of England, Wales and Northern Ireland (2003) provide guidelines on production of facial composites. The guidance states: "Where more than one image is available [from different witnesses] and it is certain that they are of the same person, they may be used in combination…" (ACPO, 2003, p. 12). Unfortunately it is not clarified whether "in combination" means that multiple images may be circulated or it refers to production of a between-witness morph,

One note of caution needs to be made. Wells Charman and Olson (2005) found that building a facial composite using a conventional software system (FACES) impaired the ability of witnesses to subsequently identify the culprit in a target-present lineup, but

did not affect the rate of mistaken identification in a target-absent lineup. It should be noted that Wells *et al.*'s data imply a loss of sensitivity but not an increased risk of mistaken identification. This would suggest that it would be poor practice for a witness who has produced a composite to attend a subsequent lineup identification procedure. If multiple witnesses are available, those who have not participated in any prior identification procedure should be invited to attend a formal identification procedure that would be used for evidential purposes. Therefore, good practice for formal eyewitness identification provides a barrier to use of between-witness morphs. If there are multiple witnesses the police may prefer to ask only one witness to construct a likeness and to retain the remaining witnesses for potential identification of a suspect after arrest. However it is not always possible to avoid use of multiple identification procedures, for example, if there is only a single witness. Therefore it would be prudent to conduct research to evaluate the effect of producing composites using a new evolution method such as EFIT-V on subsequent lineup performance. The effect of production of a single composite should be compared to production of multiple composites.

References.

Association of Chief Police Officers of England, Wales and Northern Ireland (2003) *National Working Practices in Facial Imaging.* London: ACPO

Brace, N. A., Pike, G. E. & Kemp, R. I. (2000). Investigating E-fit using famous faces. In: A. Czerederecka, T. Jaskiewicz-Obydzinska & J. Wojcikewicz (Eds.) *Forensic Psychology and law: Traditional questions and new ideas.* (pp.272-276). Krakow, Poland: Institute of Forensic Research Publishers.

Bruce, V., Ness, H., Hancock, P.J.B., Newman, C. and Rarity, J. (2002). Four heads are better than one: Combining face composites yields improvements in face likeness. *Journal of Applied Psychology , 87*, 894-902.

Davies, G. M. & Valentine, T. (2007). Facial composites: forensic utility and psychological research. In: R.C.L. Lindsay, D.F. Ross, J.D. Read & M. P. Toglia *Handbook of eyewitness psychology. Volume 2: Memory for people*. Mahwah: LEA. (pp. 59-83).

Ellis, H. Shepherd, J & Davies, G. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8*, 431-439.

Frowd, C, Bruce, V., McIntyre, A. & Hancock, P. J. B. (2007). The relative importance of external and internal features of facial composites. *British Journal of Psychology*, 98, 61-77.

Frowd, C., Bruce, V., Ness, H., Bowie, L. Paterson, J., Thomson-Bogner, C., McIntyre, A. & Hancock, P. J. B. (2007). Parallel approaches to composite production. *Ergonomics*, 50, 562-585.

Frowd, C, Bruce, V., Smith, A. J. & Hancock, P. J. B. (2008). Changing the face of criminal identification. *The Psychologist*, 21, 668-672.

Frowd, C, Bruce, V., Smith, A. J. & Hancock, P. J. B. (2008). Improving the quality of facial composites using a holistic cognitive interview. *Journal of Experimental Psychology: Applied*, 14, 276-287.

Frowd, C., Bruce, V., Ross, D., McIntyre, A. & Hancock, P. J. B. (2007). An application of caricature: How to improve the recognition of facial composites. *Visual Cognition*, 15, 954-984.

Frowd, C, Bruce, V., Plenderleith, Y. & Hancock, P. J. B. (2006). Improving target identification using pairs of composite faces constructed by the same person. *IEE conference of crime and security* (pp. 386-395). London: IET

Frowd, C., Carson, D., Ness, H., McQuiston, D., Richardson, J. Baldwin, H & Hancock, P. (2005a). Contemporary composite techniques: The impact of a forensically-relevant target delay. *Legal and Criminological Psychology, 10,* 63-81

Frowd, C., Carson, D., Ness, H., Richardson, J. Morrison, L., McLanaghan, S. & Hancock, P. J. B. (2005b). A forensically valid comparison of facial composite systems. *Psychology, Crime & Law*, 11, 33-52.

Frowd, C., Hancock, P. J. B. & Carson, D. (2004). EvoFIT: A holistic evolutionary facial imaging technique for creating composites. *Association for Computing Machinery Transactions on Applied Psychology*, *1*, 19-39.

Gibson, S. Pallares Bejarano, A. & Solomon, C. (2003). Synthesis of photographic quality facial composites using evolutionary algorithms. In: R. Harvey & J.

A. Bangham (eds.) *Proceedings of the British Machine Vision Conference 2003.* (pp. 221-230). London: British Machine Vision Association.

Gibson, S.J., Solomon, C.J., Maylin, M.I.S. & Clark, C. (2009).New methodology in facial composite construction: from theory to practice. *International Journal of Electronic Security and Digital Forensics*, 2, 156-168

Hasel, L. E. & Wells, G. L. (2007). Catching the bad guy: Morphing composite faces helps. *Law and Human Behavior, 31*, 193-207.

Leder, H. & Bruce, V. (2000). When inverted faces are recognized: The role of configural information in face recognition. *Quarterly Journal of Experimental Psychology, 53*, 513-536.

Light, L.L., Kayra-Stuart, F., Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 212-228.

Ness, H. (2005). Improving composites of faces produced by eyewitnesses . Unpublished PhD Thesis. University of Stirling.

Rhodes, G., & Tremewan, T. (1996). Averageness, exaggeration and facial attractiveness. *Psychological Science, 7,* 105-110.

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

Valentine, T., Darling, S. & Donnelly, M. (2004). Why are average faces attractive? The effect of view and averageness on the attractiveness of female faces. *Psychonomic Bulletin & Review, 11***,** 482-487.

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning & Memory, 7,*181-190.

Young, A. W., Hay, D. C., McWeeny, K. H. Flude, B. M. & Ellis, A. W. (1985). Matching familiar and unfamiliar faces from internal and external features. *Perception, 14*, 737-746.

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

Footnotes

1. EFIT-V was formerly known as EigenFIT. It is available from http://www.visionmetric.com.

2. Each participant attempted to name 10 images of one type only (e.g. 10 first-made composites, or 10 second-made). Therefore some participants, in particular those who watched one soap only, would only get to see a single best-rated image, whereas others would see a number of best-ranked images as these were not spread evenly across image-types. Therefore this analysis was restricted to participants who watched both soaps.

Annex 1

List of actors who served as the 'target' faces in experiments 1 and 2.

Actors from '*Neighbours*', an Australian soap broadcast five days per week in the UK:

Patrick Harvey (character name: 'Connor O'Neill'),

Kevin Harrington ('David Bishop'),

Stephan Lovett ('Max Hoyland'),

Stefan Dennis ('Paul Robinson'),

Blair McDonough ('Stuart Parker')

Actors from '*Eastenders*', a British soap broadcast four days per week in the UK:

Shane Richie ('Alfie Moon'),

Nigel Harman ('Dennis Rickman'),

Ricky Groves ('Gary Hobbs'),

Adam Woodyatt ('Ian Beale'),

James Alexandrou ('Martin Fowler').

List of actors who served as the 'target' faces in experiment 3 and 4.

Actors from '*Hollyoaks*', a British soap broadcast five days per week in the UK:

Stuart Manning ('Russ Owen'),

Andrew Moss ('Rhys Ashworth'),

Anthony Quinlan ('Gilly Roach')

Kent Riley ('Zak Ramsey')

Ashley Taylor-Dawson ('Darren Osborne')

Table 1

Mean ranking from Experiment 1 by participants who were familiar with the target actors.

	Between-witness morphs	Within-witness morphs	Individual composites	Best composites	First composites
From memory	N/A	8.24 (20)	11.30 (80)	9.80 (20)	11.43 (20)
From View	N/A	8.95 (20)	12.53 (80)	11.09 (20)	13.10 (20)
Mean ranking	5.78 (10)	8.59 (40)	11.92 (160)	10.44 (40)	12.27 (40)

Note: The number of ranks contributing to each mean is given in parentheses. 'Best' composite refers to the composite ranked in each experimental condition as the best by the witness who produced the composites. 'First' composite refers to the first composite produced by each witness in each condition.

Table 2

Mean similarity rating from Experiment 4 by participants who were familiar with the target actors.

		Actor 1		Actor 2		Actor 3		Actor 4		Actor 5		Mean	
		M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*	M	*SD*
Individual composites	Internal	2.71	*0.99*	4.22	*1.19*	3.52	*1.10*	3.17	*1.04*	3.83	*1.28*	3.49	*0.98*
	External	4.00	*1.15*	4.89	*1.34*	4.21	*1.22*	3.86	*1.15*	4.02	*1.05*	4.19	*0.98*
	Whole	3.15	*1.20*	4.51	*1.27*	3.54	*1.14*	3.22	*1.22*	3.51	*1.18*	3.59	*1.06*
	Mean	3.29	*1.03*	4.54	*1.16*	3.76	*1.06*	3.41	*0.99*	3.79	*1.02*	3.76	*0.96*
Within-witness morphs	Internal	3.33	*1.18*	5.13	*1.42*	4.13	*1.26*	4.08	*1.21*	4.95	*1.36*	4.33	*1.00*
	External	4.63	*1.27*	4.86	*1.23*	5.09	*1.35*	3.94	*1.03*	4.93	*1.27*	4.69	*0.92*
	Whole	4.04	*1.35*	5.24	*1.32*	4.63	*1.44*	3.71	*1.12*	4.82	*1.62*	4.49	*1.13*
	Mean	4.00	*1.11*	5.08	*1.19*	4.61	*1.14*	3.91	*0.92*	4.90	*1.19*	4.50	*0.95*
Between-witness morphs	Internal	4.05	*1.74*	6.05	*1.91*	5.20	*1.47*	4.65	*1.35*	5.45	*1.63*	5.08	*1.18*
	External	4.78	*1.62*	5.00	*1.81*	5.68	*1.42*	4.58	*1.91*	5.63	*1.93*	5.13	*1.04*
	Whole	4.70	*1.65*	6.00	*1.71*	6.33	*1.58*	5.33	*1.85*	6.00	*1.95*	5.67	*1.28*
	Mean	4.51	*1.21*	5.68	*1.55*	5.73	*1.16*	4.85	*1.36*	5.69	*1.47*	5.29	*1.04*
Mean	Internal	3.36	*1.18*	5.13	*1.38*	4.28	*1.09*	3.97	*1.05*	4.74	*1.20*	4.30	*0.99*
	External	4.47	*1.20*	4.92	*1.35*	4.99	*1.06*	4.13	*1.12*	4.86	*1.27*	4.67	*0.92*
	Whole	3.96	*1.31*	5.25	*1.30*	4.83	*1.16*	4.08	*1.23*	4.78	*1.43*	4.58	*1.10*
	Mean	3.36	*1.18*	5.13	*1.38*	4.28	*1.09*	3.97	*1.05*	4.74	*1.20*		

Figure Captions

Figure 1: Examples of individual composites and morphs created of the actor James Alexandrou (Martin Fowler from TV soap Eastenders). Panel A illustrates the equal weight given to the construction of a between-witness morph from the best self-ranked composites of four different participants. Panel B illustrates a within-witness morph created by combining four composites produced by a single participant (witness D), with greater weight given to the composites ranked better by the witness. This figure is a monochrome reproduction of color images that were used in the Experiment 1 and 2.

Figure 2 Percentage of participants in Experiment 2 shown as a function of their conditional naming rate and the type of composite. Key: BWM = Between-witness morph; WWM = Within Witness Morph; IND = Individual composite.

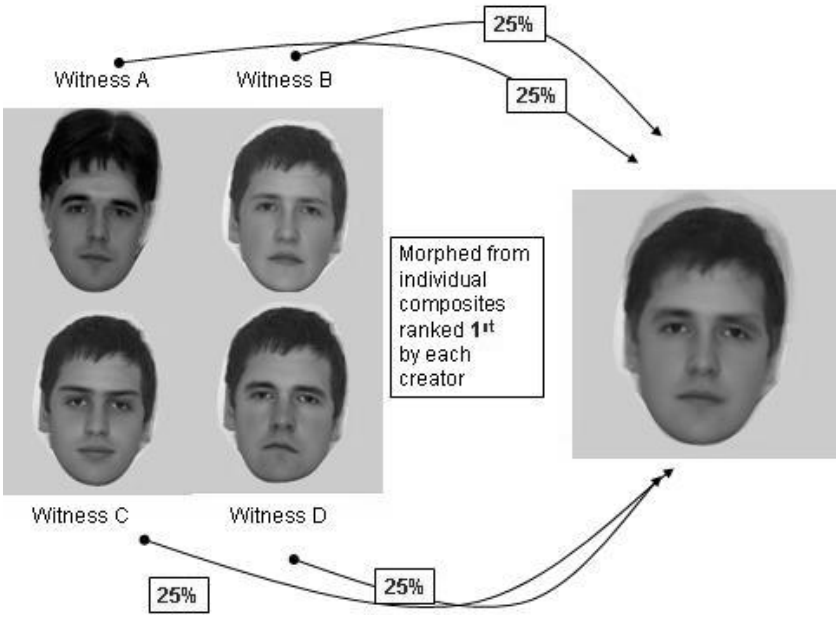
Figure 3: Mean image similarity rating as a function of facial type and actor (error bars denote standard error of the mean) in Experiment 3.

Figure 4: Whole, external and internal images of one of the targets (Nick Pickard).

Figure 5: Mean image similarity rating as a function of facial type and image presentation (error bars denote standard error of the mean). This figure is a monochrome reproduction of color images that were used in Experiment 4.

Figure1

Panel A



Panel B

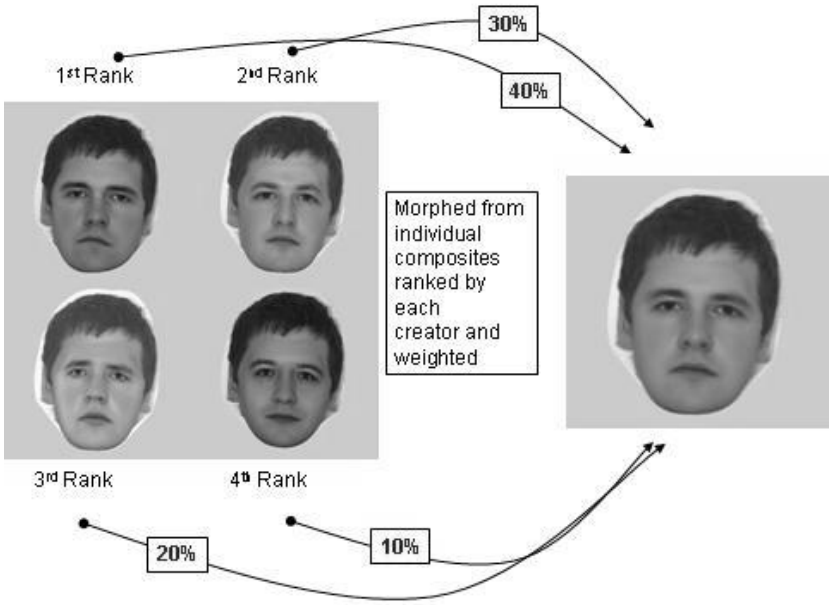


Figure 2

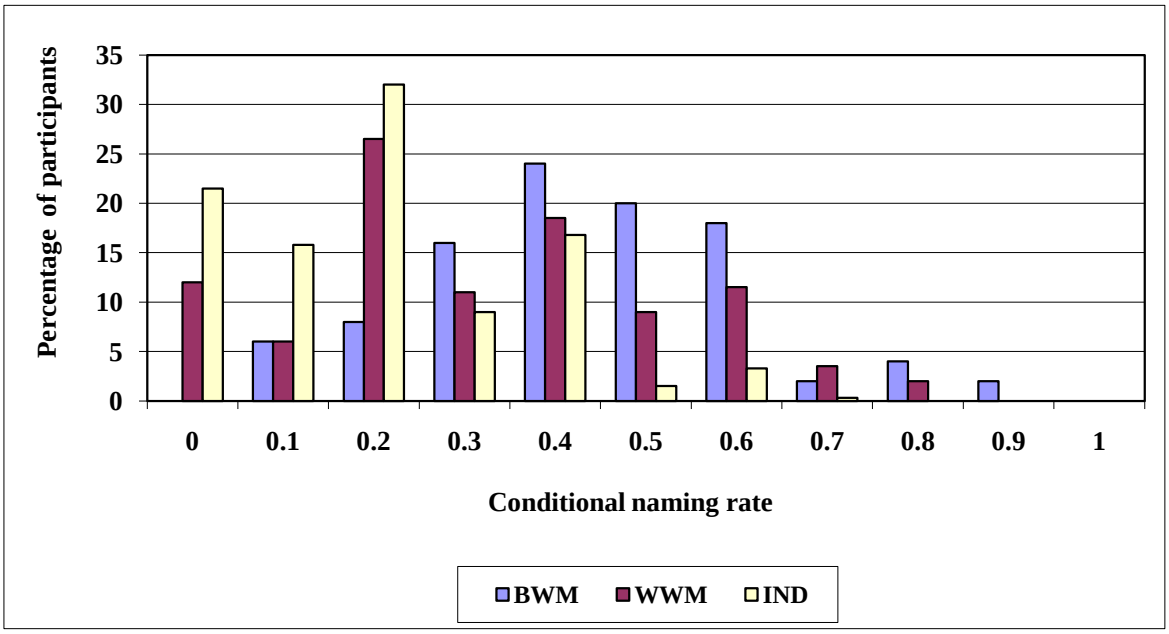

Percentage of participants
35
30
25
20
15
10
5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Conditional naming rate
BWM
WWM
IND

Figure 3

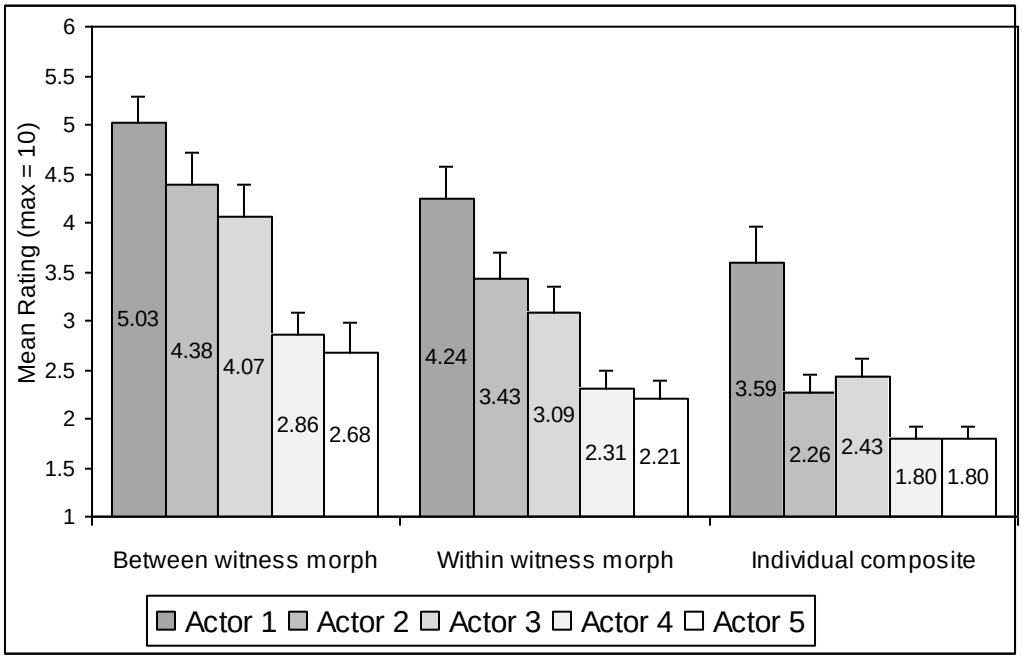
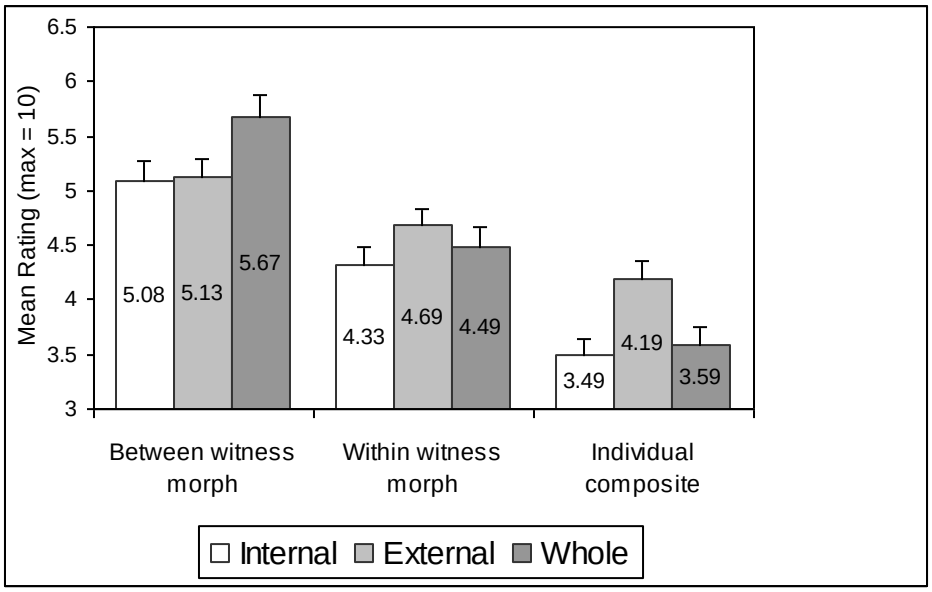

Mean Rating (max = 10)
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
5.03
4.38
4.07
2.86
2.68
4.24
3.43
3.09
2.31
2.21
3.59
2.26
2.43
1.80
1.80
Between witness morph
Within witness morph
Individual composite
Actor 1
Actor 2
Actor 3
Actor 4
Actor 5

Figure 4

Whole External Internal

Figure 5


Mean Rating (max = 10)
6.5
6
5.5
5
4.5
4
3.5
3
5.08
5.13
5.67
4.33
4.69
4.49
3.49
4.19
3.59
Between witness morph
Within witness morph
Individual composite
Internal
External
Whole