Four heads are better than one: combining face composites yields improvements in face likeness

Vicki Bruce, Hayley Ness, Peter J.B. Hancock, Craig Newman, Jenny Rarity.

University of Stirling,

Abstract.

Four different participants constructed face composites, using “PRO-Fit”, of familiar and unfamiliar targets, with reference images present or from memory. The "mean" of all four composites, created by morphing (4-Morph) was rated as a better likeness than individual composites on average, and was as good as the best individual likeness. When participants attempted to identify targets from line-ups, 4-Morphs again performed as well as the best individual composite. In a second experiment participants familiar with target women attempted to identify composites, and the trend showed better recognition from multiple composites, whether combined or shown together. In a line-up task with unfamiliar participants, 4-Morphs produced most correct choices, and fewest false positives from target absent or target present arrays. These results have practical implications for the way evidence from different witnesses is used in police investigations.

Introduction

Eyewitnesses to crimes often produce crucial and very influential evidence about the identity of the criminal, which can have a major impact upon criminal conviction (Wells, 1993; Wells et al, 1995). Often police will collect evidence of facial appearance from a witness by inviting them to attempt to build a composite image of the face of the person seen committing a crime. Manual composite systems which have been used extensively, at least until recently, in the US (Identikit) and Europe (Photo-fit) invite the witness to select a set of facial features corresponding to those remembered of the criminal, and to compose a whole face from these individual components. Occasionally, striking likenesses of convicted criminals have been built using such systems (e.g. among many composites produced by witnesses to crimes which were eventually associated with Peter Sutcliffe, the "Yorkshire Ripper", a small number bore a close resemblance to Sutcliffe). However, formal psychological evaluation of such systems tended to reveal that they were rather poor at eliciting good likenesses (e.g. Ellis, Shepherd & Davies, 1978; Laughery & Fowler, 1980).

Over the past ten years or so, manual systems have largely been superseded by electronic composite systems. These allow more flexibility in the placement of features, and better blending of different features to remove the lines which marred the appearance of Photofit composites. Nonetheless, electronic composite systems, whether based upon line-drawn features ("Mac-a-Mug") or photographed ones ("E-Fit" and "CD-Fit" - two very similar systems in use in the UK), ultimately still require the witness to recall individual features of remembered faces. This process is inherently difficult and perhaps even incompatible with the way the human brain codes faces for its more regular task of recognition. Wells & Hryciw (1984) for example, found instructions which promoted good recognition of faces impeded composite reconstruction, and vice versa. Recent evaluations of the Mac-a-Mug electronic composite system have found it yields poor likenesses when composites are to be constructed from memory (Koehn & Fisher, 1997; Kovera

et al, 1997, Wogalter & Marwitz, 1991). In a recent comparison of Photo-fit with E-fit, Davies, Van der Willik and Morrison (2000) found that E-fit was superior to Photo-fit only when witnesses attempted to produce copies of faces which were physically present for construction (revealing the superiority of the E-Fit system in terms of feature range and flexibility). When faces were constructed from memory the systems produced similar, and disappointingly poor likenesses – perhaps indicating that the limitation for composite systems used in forensic contexts lies in the task required of the witness, rather than in limitations of the system. So, although we now have a set of improved computer-based tools with which to build images of remembered faces, the process of recalling faces is inherently difficult, and any individual attempt to generate a composite is likely to contain some distortion of the true appearance.

In this paper we investigate whether better likenesses might be produced if multiple witness composites of the same remembered person were generated and used in combination. Bennett, Brace, Pike and Kemp (1999) reported an investigation where identification of people from presented composites was enhanced when several different composites were shown, compared with when individual composites were shown. Our aim, however, was to explore whether a single composite's effectiveness could be enhanced if it was built from the combined memories obtained from several independent witnesses. McNeil et al (1987) attempted something similar by building composites based upon the most frequently chosen Identikit features from groups of eight independent witnesses. 'Modal' composites were given higher similarity ratings than individual composites in their study. However, this approach requires relatively large numbers of witnesses and in any case cannot easily be adapted to electronic systems where feature locations as well as types are varied. Our own approach was based upon simple merging of the images themselves.

Our reasoning was that since deviations from "true" appearance in individual composites are unlikely to be correlated, merging the different individual composites together should tend to reinforce correct aspects and de-emphasise incorrect ones. Essentially, we are exploring the notion that each individual recall attempt is some deviation from the 'prototype' (veridical) image of the target person. Superimposing these different recall attempts should reveal an image closer to this prototype. Research into the prototype effect shows that participants tend to find an unpresented prototypical pattern (Posner & Keele, 1968) or face (Solso & McCarthy, 1981; Bruce, Doyle, Dench & Burton, 1991; Cabeza, Bruce, Kato & Oda, 1999) at least as familiar as items which have actually been studied. In the discussion of this article we will return to this phenomenon when considering the potential for implementation of the reported effects.

So, the aim of the paper was to explore whether combining individual composite images by "morphing" produced better likenesses. We conducted two experiments – the first used rather artificial methods in order to examine whether there was any promise in the general approach. Experiment 2 used a methodology which much more closely simulated the kind of events involved in actual eyewitness memory and composite production and usage.

Experiment 1:

In the first stage of the experiment, composites were generated of famous and of unfamiliar faces, either with targets present, or from memory, such that four different composites were produced in each condition. In the second stage, individual and morphed composites were rated for likeness against their intended targets. In the third stage, the best, worst, morphed and multiple (all four) composites were shown alongside line-ups of six alternative faces, and participants asked to choose which face matched the composite(s). For clarity, we here describe the procedure and results of each stage in sequence.

Stage 1: Generating individual composites and morphs.

Method

Participants

Thirty-two adults drawn from a wide range of ages, backgrounds and types of employment were paid £10 for a session lasting up to 2 hours.

Materials

Four unfamiliar and four famous male targets were chosen from a larger set by someone unconnected with composite construction, so that the operators in the study remained blind to the particular targets chosen. The set of famous faces from which the four targets were drawn were all people known from other experimental work in our laboratory to be familiar to a wide range of participants drawn from the general public. Criteria for selection for this task were that the targets should be clean-shaven males, lacking in major distinguishing features that would lead the operators to become aware of whose face they were building. The four famous targets selected were the American actors George Clooney and Tom Hanks, the British singer Robbie Williams and the Scottish actor Robert Carlyle (all very well-known in the UK). The four unfamiliar faces were selected from a casting agency catalogue of photographs of models and actors (so that the type of face shown, in terms of pose, was similar to famous faces) to cover a similar age range to the famous face targets. Composites were generated using '*Pro*fit' (the new trade name for CD-Fit in the UK) on a Hi-Grade Ultis PV3 Personal Computer. The composites were morphed using the software package 'Sierra Morph v. 2.5'.

Design

Composites were created of faces which were familiar (famous) or unfamiliar, and the composites were created with a reference image present, or from memory. Each participant constructed one composite with reference image present, to familiarise the composite system, and then a second

composite of a different target from memory, for which a longer time was allowed. Half the participants constructed a familiar face first and half constructed an unfamiliar face first. The result was that eight independent composites, from different participants, were formed of each face - four with reference present and four from memory. Experimenters remained blind to the targets chosen until after Stage 1 was completed so that they could not influence the features selected by the participants. [i]

Procedure:

1st composite: Face Present

Participants were given a photograph of a face (familiar or unfamiliar) and were asked to describe this face to the experimenter (who could not see the picture). The procedure was identical irrespective of the familiarity of the target, but participants were asked to avoid mention of personal identity in their descriptions. The experimenter attempted to use Pro-Fit to recreate the face under the instruction of the participant, until the participant was satisfied with the result within a time limit of 30 minutes.

2nd composite: Memory

Participants were asked to study a face for 30 seconds (either of a familiar or unfamiliar target: the opposite type to that used for the first composite). Showing the photograph of familiar targets ensured that in all conditions of the experiment there was equal influence from the specific photographs chosen. The participant then attempted to work with the experimenter to build a composite from memory, following recommended procedures for composite production using Pro-Fit in police contexts. Cognitive interview techniques (Fisher & Geiselman, 1992) were used in which the witness was first asked to describe the face as fully as possible, as if to a friend, then probed to recall information about specific features. They were then encouraged to use imagery and to describe the face again, and they were offered pen and paper to supplement descriptions with sketching if they so wished. Only after this stage did the witness and operator proceed to

construction of a composite. The process was terminated when the witness expressed satisfaction with the result, or at the end of a time limit of 90 minutes had elapsed,

After each composite was completed, participants gave an estimate of the rated similarity of their completed composites to the original face on a scale of 1-10 (with 10 being most similar). Participants' expressions of confidence in the goodness of their composites were used to weight the Morph combinations at the next phase.

Morphing

The above procedure generated four original composites for each target in each condition (reference present and memory). The first two of these were morphed together by placing markers on the main features on the first face (round the eyes, eyebrows, nose, mouth, hairline, jaw-line and ears) and matching each marker to the corresponding area on the second face. A morph was then produced of 50 percent of each face. This was then repeated with the next two images from the same set. The morphs created from pairs of faces are termed '2-Morphs'. The pair of 2-Morphs were in turn morphed together to produce a morph containing 25 percent of each face (Figure 1 shows four original composites alongside a 4-Morph created in this way).

Figure 1 about here

For weighted morphs, the percentage given to each composite depended upon the confidence rating given by its creator. For example, if the first composite gained a confidence rating of 4 and the second a confidence rating of 6 then the weighted 2-Morph of these two composites comprised 40% of composite 1 and 60% of composite 2.

As a result of Stage 1, we had for each target, for each condition (reference present or absent) a set of four individual composites, four 2-Morphs (two weighted, two unweighted) and two 4-Morphs (one weighted, one unweighted) – ten composites in total per target per condition.

Stage 2: Likeness Ratings

Method

Participants

Forty unpaid participants (male and female) were approached and recruited individually from students attending an Open University summer school at the University of Stirling.

Materials and Design

Booklets were created in which each sheet displayed one composite printed beside the photograph of the corresponding target face. One booklet showed famous targets, the second showed the unfamiliar ones.

Similarity ratings of the famous and unfamiliar composites were collected from different groups of participants. For each target face, in each condition, there were ten different composites (four individual, four 2-Morphs, and two 4-Morphs). Type of composite (individual, 4-Morph-Weighted/Unweighted; 2-Morph Weighted/Unweighted) and whether produced Present or from Memory were both within-subjects and within-items factors. Thus each booklet contained 80 sheets (reference present/absent x four targets x ten different types of composite), and their order was randomised.

Procedure

Participants were shown either the familiar or unfamiliar folder, and were told that each composite represented an attempt by the witness to construct a recognisable image of the target

person. They were asked to rate how good each attempt was by rating how closely each composite resembled the face alongside it on a scale of 1 (no resemblance whatsoever) to 10 (maximum resemblance). The experimenter recorded ratings manually.

Results

Figure 2 and Table 1 about here please

Figure 2 summarises the mean similarity ratings across the different types of composite, for simplicity averaging over the ratings for weighted and unweighted composites which did not differ in our statistical analyses. Across conditions, the individual composites on average were rated as least similar, with the 2-Morphs performing better, and the 4–Morphs performing best. Table 1 provides all cell means and standard deviations.

These data were analysed using a 2 (famous/unfamiliar target) x 2 (reference present/memory) x 5 (composite type: individual; 2-Morph weighted; 2-Morph unweighted; 4-Morph weighted ; 4-Morph unweighted) ANOVA by subjects (F1) and by items (F2). The first factor was between-subjects and between-items; the second factor was between subjects but within-items; the third factor was within-subjects and within-items. The overall difference between the five composite types was significant, $F1 (4, 152)=89.8$, $F2 (4, 24) = 52.12$, $p<.001$ (calculated effect size was large using Cohen's (1969) definition on both subjects ($f=0.39$) and items ($f=0.56$) analysis). There was also a significant interaction between familiarity and composite type ($F1(4, 152)= 7.99$, $F2 (4, 24) = 4.67$, $p<0.01$, see Figure 2). Figure 2 shows that the similarity ratings for the familiar composites are somewhat lower (in all conditions) than the unfamiliar composite ratings. The data trends for the composite types appear similar for each familiarity type with the original composite performing lowest, the 2-Morphs performing higher and the 4-Morphs being rated as

most similar. However these trends appear stronger for the unfamiliar faces. Simple main effects analyses followed by Scheffe tests (on both subjects and items data) indicated that for both familiar and unfamiliar faces, the original composites gained the lowest similarity ratings; but only for the unfamiliar faces were there significant differences between the 4-morphs (both weighted and unweighted, which did not differ) and the 2-morphs (weighted and unweighted did not differ).

Composites created with the target face present were rated as more similar than composites created from memory ($F1(1, 38) = 137.5$, $p<0.001$, but $F2(1,6) = 5.18$, $p=0.06$ suggesting that the effect was not so consistent across different items, though note that the small size of the item set reduces power in this analysis;), means were 4.0 (S.D.1.0) from memory and 4.9 (S.D. 1.1) when target faces were present (calculated effect size by items $f=0.42$, which is again large). No other effects were shown in both subjects and items analyses.

By looking at the data by-items it was possible also to identify for each item, which of the four composites achieved the best and which the worst overall similarity ratings. We then repeated the ANOVA on the items data to compare each of the morph conditions with the best single composite obtained for each item. There was still a main effect of composite type ($F(4,24) = 7.96$, $p<0.01$) – reflecting the overall better likeness ratings given to the 4-Morphs (mean 5.0, S.D. 1.2 averaged over weighted and unweighted) compared with the 2-Morphs (mean 4.4, S.D. 1.0, averaged across weighted and unweighted) and best single composites (mean 4.4), S.D. 1.0) which did not differ.

Stage 3: Identification via line-ups.

Method

Participants

Forty unpaid participants were recruited from local universities where they were approached in public areas to assist with a study on eyewitness memory.

Materials and Design

A line-up of six faces of similar hairstyle and approximate age (famous for famous targets, unfamiliar for unfamiliar targets) was created for each individual target, and target position within the line-up was varied across different arrays.

The best and worst individual composites were selected for each target on the basis of the Stage 2 ratings. These were used to probe recognition from line-ups, and performance was compared with that obtained using the (unweighted) 4-Morph, and that obtained from the full set all four individual composites. Different groups of participants saw composites from target-present and memory generation conditions, but other factors were varied within participants, so that each participant saw each of the eight targets in a different composite condition.

Procedure

Participants were shown a line-up of six individual photographs together with either the best, the worst, the 4-Morph, or all four individual composites. They were told that the composite(s) were attempts to generate a likeness of one of the six faces, and asked to choose which of the six they thought it was. Each participant attempted one line-up for each individual familiar and unfamiliar target, seeing a different one of the four composite conditions for each familiar target, and for each unfamiliar target. Allocation of targets to composite conditions was rotated so that a total of five subjects saw each target in each composite condition.

Results

Figure 3 about here please

The number of times (max 5) each target was correctly picked from the 6-alternative line-ups is summarised in Figure 3. A 2 (familiarity) x 2 (reference present/memory) x 4 (composite type) mixed design by-items ANOVA showed significant effects of familiarity ($F(1, 6)=15.70$, $p<0.01$) with more correct choices for familiar targets, and type of composite ($F(3,18) =4.67$, $p<0.05$). Overall, the 4-Morph condition was significantly better than the worst ($p<0.05$ Scheffe and t-test) and marginally better than the multiple condition ($p<0.05$ t-test, but $p<0.1$ Scheffe), and the multiple condition was also marginally better than the worst ($p<0.05$ t-test; $p<0.1$ Scheffe test). The calculated effect size for the main effect of composite type was again large ($f=0.61$). However, a three-way interaction between familiarity, composite type and whether the targets were produced from memory or present was also significant ($F(3, 18) = 3.63$, $p<0.05$). The general superiority of the 4-Morph condition was not sustained across all four familiarity-target present/memory combinations. In particular, in the condition where unfamiliar faces are generated from memory, target choices are infrequent and there were no significant differences across the different types of composite (see Fig 3).

Discussion

Experiment 1 showed promising effects in line with our predictions – a morph of four individual composites is rated as a better likeness, and yields somewhat better recognition, than the individual composites themselves. However, Experiment 1 used a methodology with little ecological validity, since the composites were built of seen or remembered pictures rather than people. Experiment 2 attempted a replication using a method with more applied validity.

Experiment 2

In this experiment we attempted a replication of the effects of morphing, using a new set of (female) targets, and a more ecologically valid procedure. In stage 1, participants viewed a video of an unfamiliar target woman, before attempting to construct a composite of her face from memory. In stage 2, generated composites were rated for likeness by participants unfamiliar with the targets, but in stage 3, participants who should be familiar with the target women attempted to identify the composites. Finally, in stage 4, participants unfamiliar with the targets attempted to compare composites with 6-item line-ups. This time, unlike Experiment 1, half the line-ups had no target present, so that we could examine whether the different conditions were more or less likely to trigger false recognitions as well as hits.

In Stage 1 of this experiment participants attempted to create composite images in full-face (standard) and also in ¾ view, using a new prototype version of PRO-Fit. Here we are only interested in evaluations of the full-face composites generated - the ¾ view research will be reported separately.

Stage 1: Generation of composites.

Materials

The target women selected were all members of staff (teaching and/or administrative) at the University of Stirling psychology department. Composites were generated using PRO-Fit and morphs created using Sierra Morph (version 2.5).

Participants:

16 volunteers aged 18-40 years from Queen Margaret University College in Edinburgh, and unfamiliar with the women whose faces were shown, were each paid £10 for participation.

Design and Procedure

Participants were shown a 30-second video in which the target rotated their head in front of the camera and smiled and spoke. Using cognitive interview techniques, they then attempted to recall her so that the experimenter could build a composite, first in one viewpoint (FF or ¾) and then in the other. Half the participants constructed the full-face composite first and half the ¾ view. Procedure otherwise was as in experiment 1. A total time of 2 hours was allowed for the construction of the two composites.

Using the same Morph procedure as in Experiment 1 (unweighted), a 4-Morph was constructed by combining the full-face composites of each of the four target faces. Figure 3 shows an example of the four individual items, and the 4-Morph produced of one of the targets in this experiment.

Figure 4 about here please

Stage 2: Likeness ratings

Method

Twenty participants aged 17- 60 years were recruited from staff members at Tesco stores, Edinburgh to rate the composites.

Four different full-face composites were generated of each target. These individual composites and the 4-Morph resulting from their combination were rated for likeness against photographs of the target women.

Participants were shown 5 full-face composites (4 individual and one 4-Morph) of each of four targets, in a total of 20 trials. On each trial the composite was shown alongside both a full-face and ¾ portrait of the woman, and likeness of the composite to the depicted target was rated on a scale from 1 (low) to 10 (high), using the same instructions as in Experiment 1.

Results

From Stage 2, the 4-Morphs (mean likeness rating 5.4, SD =2.0) were given higher likeness ratings than average of the individual composites (mean =4.25, SD=1.5). A 4(targets) x 2 (type of composite) ANOVA[ii] revealed significant main effects of target ($F(3,57) = 36.7$, $p<0,001$), composite type ($F(1,19) = 38.06$, $p<0.001$) and their interaction. ($F(3,57)= 7.22$, $p<0.01$). The effect size for the main effect of composite size was calculated as f=0.33, medium to large according to Cohen (1966). Simple main effects analysis revealed that the mean likeness rating was higher for the morphed than individual composites for all four targets, and significantly so for two of the four targets. For each target it was possible to identify the best and the worst of the individual composites to compare with the 4-Morph. Means for best were 5.5 (SD 1.1); for the 4-Morph 5.4 (SD 2.0), and for the Worst 2.3 (SD 1.5). Thus the 4-Morph was rated as well as the best individual composite of each woman, and significantly better than the average individual composite.

Stage 3: Identification task.

Method

Participants

Thirty-two participants attempted to recognise the targets. These were 30 senior students and 2 members of staff at the psychology department, University of Stirling. Only participants who (on debriefing) indicated familiarity with all four of the target women were retained in the experiment.

Design and procedure

Participants were approached and asked to attempt to identify the composites from members of the psychology department. Each participant saw one target in the best, the worst, the 4-Morph or as a set of all four individual composites and attempted to recognise them. Allocation of targets to conditions was rotated such that a total of eight participants attempted to identify each woman in each condition.

Results and Discussion:

The recognition rates overall were relatively low and did not differ significantly between the four conditions. The highest rate was found for the condition where all four composites were shown (38% correct identifications and no false positives), followed by the 4-Morph (28% correct, 6% false positives) and best individual (22% correct, no false positives) with the worst composite yielding the lowest rate (16% correct and 6% false positive). Thus the trend is clearly in line with earlier observations – the 4-Morph appears at least as good as the best individual, though the trend in this part of the study favours multiple composites (cf. Bennett et al, 1999). False positive rates were actually very low (a 6% rate represents just two such errors) and occurred largely to one face who was misidentified as a female professor in the department who did not appear in the target set.

Stage 4: Identification from line-ups.

Participants

Participants were 64 volunteers unfamiliar with the target faces (of necessity, since Stage 2 had exhausted the pool of those familiar), recruited from students of the Open University attending a summer school at Stirling. They were aged between 17 and 60 years and had normal or corrected

to normal vision. At debriefing it was established that all subjects were unfamiliar with all faces used in the experiment.

Design and Procedure.

The same four conditions (4-Morph, best, worst and all four individual composites) were compared, in a task which required matching composite(s) against an array of female faces of similar general appearance. On half the arrays no target image was present. Thus there were eight different conditions (four types of composite, target present or absent for each), for each of the four target faces. Each individual participant attempted to recognise each target just once, in one type of composite, with half their trials using a present array and half using an absent array. Eight different books of arrays were created so that each target face was attempted in each condition by a total of eight participants.

For each female target a set of distractors was selected from verbal descriptions given in the construction phase. These were matched for hairstyle, face shape and approximate age. The distractors plus target (or distractors alone for target absent arrays) were presented in monochrome arrays in two rows of three faces, with target position varied from array to array within target-present arrays.

Each participant was given a brief description of the ways in which the composites had been constructed, being informed that the composites were constructed from memory and represented a likeness of the actual person and not an identical image. They were told to compare the composite (or set of four attempts at the same target) against the array of faces, and that the target person ‘may or may not be’ present in the array. They were asked to decide whether they thought the target person was there and if so, which one it was. No time limit was imposed and the same procedure was used for all four arrays in the booklet.

Results.

Table 2 about here please.

Table 2 summarises performance at Stage 4. The 4-Morph condition yields the most correct responses from both target present and target absent arrays (mean 41% correct), and generates the fewest false choices (mean 48%) from either type of array. Here, the trend is for the four-composite condition to yield the poorest performance (mean 21% correct, 75% false choices) followed closely by the worst individual (28% correct; 68% false choices) and the best individual (27% correct; 66% false choices) which appear virtually the same. Rates of correct responses to the four conditions of the experiment did not differ significantly when all four conditions were compared using Chi-squared. However, a comparison between the 4-Morph condition and the average of the other three conditions was significant (Chi-squared = 4.68, $p<0.05$).

Discussion

In Stage 3 of this experiment showing all four composites together appeared to be marginally the best condition, whereas this was not shown with the 6-alternative FC line-ups in Stage 4, nor in Expt 1, where the four composite condition appeared substantially worse. This difference may reflect task demands. When the task is to identify who a person is then several different variants may be useful (since one may trigger recognition where others fail). But where the task is to compare with other items in a line-up the variation may create confusing false matches with distractors. However, it is important to stress that there was no significant advantage for the multiple condition over the 4-Morph at Stage 3 of this experiment, and again this experiment shows that the 4-Morph condition fares well in both identification and matching tasks.

The relative superiority of the 4-Morph condition is further bolstered by another experiment recently completed in our laboratory by Pirie (2001) using a similar design to Experiment 2 with students attempting to recognise composite images of their classmates. Composites had been constructed from memory (but of studied pictures rather than video) by witnesses unfamiliar with the images. Pirie compared recognition rates with a single composite (selected at random from the four available for each face), a 4-morph, and all four composites, and these conditions yielded rates of 17%, 56% and 42% respectively. There was no significant difference between the morph and multi-composite conditions, but both were better than the single composite condition. Note in this case, the non-significant trend in an identification task favoured the morph rather than multiple composite condition. On the basis of these results, it seems safe to conclude that the trends for identification favour the use of evidence from multiple witnesses – whether separately or combined – over individual witness composites.

The low rates of identification obtained overall in Stage 3 of Experiment 2 here are in line with other research using composite systems (Davies et al, 2000, Koehn & Fisher, 1997; Kovera, Penrod, Pappas & Thill, 1997). The low rates at Stage 3 may seem surprising given that our participants knew that they were attempting to identify female members of the psychology department. The occasional false positives suggest that some guessing may have gone on. However, the pool of individual women within the department whose identities could have been probed is actually quite large (there are 35 or more in total, and at least 15 who would be well known to most undergraduates), making guessing an unhelpful strategy. In Stage 4, like Experiment 1, we found that the 4-Morph gave rise to higher rates of choosing from line-ups than did the single composite comparisons. Additionally, this condition seems also to protect rather than encourage false choices when no target is present, showing that its effects arise through changes in discriminability rather than bias.

General Discussion

Both experiments show that better likenesses are obtained when facial images from multiple witnesses are combined. 4-Morphs, and less strongly 2-Morphs (Expt 1) were clearly better than the average individual composite and were at least as good as the best individual composite (better in Experiment 1, and the same in Experiment 2). Given that there is no way of knowing whether an individual witness will produce a composite which is as good as the best that could be obtained from several witnesses, it seems that the use of multiple witnesses and composites is likely to be advisable.

What have we learned about the process of recalling faces from memory using this research? It appears that individual witness recollections may indeed be considered as variations from the ideal or 'prototype' (Posner & Keele, 1969). A combination of their memories reveals a better impression of the 'prototype' since it will weight accurate over inaccurate details, assuming that different witness errors are uncorrelated. The 4-Morph thus provides a better probe for similarity and recognition than many of the individual composites – and seems at least as good as the best.

The efficacy of the 4-Morph is in some respects surprising, since the combination of several images in this way should tend to reinforce typical rather than distinctive features of a face, producing a morphed composite which is more 'average' in appearance than any of the contributing composites. In current work in our laboratory we are investigating the additional effect of caricaturing the resulting 4-Morph to enhance its distinctive features, and we expect to find that a 4-Morph which has been subjected to a moderate degree of caricature will serve as an even better cue to identity. Performance using the multiple composite condition in our identification task (Stage 3, Experiment 2) also lends weight to other observations by Bennett et al (1999) that using multiple witnesses may be beneficial even if individual attempts are not combined, but presented simultaneously. Clearly there is merit in obtaining and using more than

one composite image when multiple witnesses view the same crime, even if these are never combined.

While the collection and combination of multiple witness composites appears promising, evidential requirements appear currently to preclude the combination of witness memories in this way, at least in the UK. When composites are generated in the UK, the "PACE" rules of evidence (Police and Criminal Evidence Act, 1984) , and less formal facial identification guidelines (ACPO, 1999), between them prohibit the generation of composites by more than a single witness to the same crime. (Where multiple composites have been produced, as in the case of the Yorkshire Ripper, these have been obtained from witnesses of separate individual crimes, later linked by the same serial offender.) A composite may be used in the investigatory process but it may also be used in courts of law. The ACPO (1999) booklet states that "…the composite image is therefore intended as an aid to the investigation of crime *together with provision of corroborative evidence*" (emphasis added). Composite images produced must be signed and sealed into envelopes marked with their date of production, in case they are needed in courts of law, and a warning is attached which states "to alter, add, tint, colour or change any details within the pictorial statement would amount to tampering with evidence". When there is more than one witness to a crime their likely efficacy is assessed using such factors as time of exposure to the face, and a single witness is chosen to generate a composite. "Each witness should be assessed individually on their level of recall to produce a composite image. One image should then be produced using a witness with good recall. The remainder of the witnesses can then be kept for other formal means of visual identification." And "The production of a composite image with multiple witnesses must not be attempted, as this will amount to cross contamination." Clearly this direction precludes witnesses working together to produce a combined composite. However, it is not clear that it necessarily precludes the combination of their evidence collected separately. There are clearly different issues here. One is to do with 'saving' some witnesses for other

identification tests such as line-ups or photo-spreads. Clearly these witnesses cannot build a composite after they have seen a line-up, and they probably should not go to a line-up after building a composite, since choices in a line-up or photo-spread could be influenced by incorrect features in their composite (e.g. see Jenkins & Davies, 1985). However, even if there were enough witnesses to save some for a line-up and still have several potentially available to build composites, the issue of cross-contamination might be avoided if we clearly separate out the two different uses of a composite.

The production of a 'best' composite to trigger identification can be distinguished from the use of individual composites as corroborative evidence. Even if the evidence from two or more witnesses is to be used in combination, an individual witness composite could still be independently constructed, and signed, sealed, dated and used (if needed) in a court of law. There seems no reason why individual composites created, and protected, in this way could not be merged to present a likeness on the TV or newspapers. Professionals in this area tend to agree that the principal aim of a composite is as a trigger to recognition, to assist in the investigatory process, rather than as something which is evidentially relevant in court. While protecting the record of an individual's testimony at the time a composite was produced, we should still aim to facilitate best practice in the usage of composites as clues to identity in the investigatory process.

References

ACPO/ACPOS – National working Practices in Facial Imaging. Manual produced by the ACPO Working Group for Facial Identification.

Bennett, P. , Brace, N., Pike, G. & Kemp, R. (1999). Making the most of E-Fits: Are multiple witnesses more accurate than a single witness? Paper presented at the 1st joint conference of the American Psychology-Law Society and the European Association of Psychology and Law, Dublin.

Bruce, V., Doyle, T., Dench, N. & Burton, M. (1991). Remembering facial configurations. Cognition, 38, 108-144.

Cabeza, R., Bruce, V., Kato, T. & Oda, M. (1999). The prototype effect in face recognition: Extension and limits. Memory & Cognition, 27, 139-151.

Cohen, J. (1969). Statistical power analysis for the behavioural sciences. New York: Academic Press.

Davies G, van der Willik P, Morrison LJ (2000). Facial composite production: A comparison of mechanical and computer-driven systems. Journal of Applied Psychology, 85: (1) 119-124.

Ellis, H.D. Shepherd, J.W. & Davies, G.M. (1975). An investigation of the use of the Photofit system for recalling faces. British Journal of Psychology, 66, 29-37.

Fisher, R.P. & Geiselman, R.E. (1992). Memory enhancing techniques for investigative interviewing: The cognitive interview. Springfield, IL: Thomas.

Jenkins, F. & Davies, G. (1985). Contamination of facial memory through exposure to misleading composite pictures. Journal of Applied Psychology, 70, 164-176.

Koehn, C.E. & Fisher, R.P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. Psychology, Crime and Law, 3, 209-218.

Kovera, M.B., Penrod, S., Pappas, C. & Thill, D. (1997). Identification of computer-generated facial composites. Journal of Applied Psychology, 82, 235-246.

Laughery, K.R. & Fowler, R.F. (1980). Sketch artist and Identi-Kit procedures for recalling faces. Journal of Applied Psychology, 65, 307-316.

McNeil, J.E., Wray, J.L., Hibler, N.S., Foster, W.D., Rhyne, C.E. & Thibault, R. (1987). Hypnosis and Identi-Kit: A study to determine the effect of using hypnosis in conjunction with the making of Identi-Kit composites. Journal of Police Science and Administration, 15, 63-67.

Pirie, S. (2001). Recognising familiar faces from computer-drive composites: A comparison of composites created using single and multiple witnesses. Final year undergraduate honours dissertation, University of Stirling.

Posner, M.I. & Keele, S.W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 83, 304-308.

Solso, R.L. & McCarthy, J.E. (1981). Prototype formation of faces: A case study of pseud-memory. British Journal of Psychology, 72, 499-503.

Wells, G. & Hryciw, B. (1984). Memory for faces: Encoding and retrieval operations. Memory & Cognition, 12, 338-344.

Wells, G.L., Small, M., Penrod, S., Malpass, R., Fulero, S.M. & Brimacombe, C.A.E. (1995). Eyewitness identification procedures: Recommendations for lineups and photospreads. Law and Human Behaviour, 22, 603-647.

Wogalter, M.S. & Marwitz, D.B. (1991). Face composite construction – in-view and from memory quality and improvement with practice. Ergonomics, 34, 459-468.

Author Note

Vicki Bruce, Hayley Ness, Peter J.B. Hancock, Craig Newman, Jenny Rarity, Department of Psychology, University of Stirling.

This research was funded by an EPSRC/DTI LINK project to the University of Stirling and ABM-UK Ltd. We thank Dr Leslie Bowie of ABM-UK for provision of *Pro*Fit software used in these experiments, Mike Burton and two anonymous referees for comments on an earlier draft of the manuscript. Krista Overvliet assisted with the data collection for Experiment 2. The design of Experiment 2 was influenced strongly by suggestions made by Richard Kemp about appropriate methods to test the efficacy of facial composite systems.

Correspondence concerning this article should be addressed to Vicki Bruce, Department of Psychology, University of Stirling, Stirling FK9 4LA, Scotland, UK. Electronic mail may be sent via Internet to Vicki.Bruce@stir.ac.uk

Figure legends

Figure 1: Four individual composites, and the 4-Morph (below), of actor George Clooney.

Figure 2: Likeness ratings from experiment 1 (on scale from 1 to 10).

Figure 3: Identification from six-alternative line-ups (average score per target, max = 5).

Figure 4: Four individual composites of one of the female targets used in Experiment 2, and the 4-Morph (below) alongside a picture of the target woman.

Table 1: Means and standard deviations for each condition of Experiment 1.

Composite:	Single		2-Morph		2-MorphW		4-Morph		4-MorphW	
	Mem	Pres	Mem	Pres	Mem	Pres	Mem	Pres	Mem	Pres
Fam										
M	3.1	3.7	3.9	4.3	3.8	4.7	4.2	4.8	4.0	5.0
SD	1.0	1.3	1.5	1.5	1.4	1.4	1.6	1.3	1.6	1.7
Unf										
M	3.0	4.0	3.9	5.2	4.0	5.1	5.0	5.9	4.8	6.3
SD	1.1	1.4	1.2	1.3	1.4	1.4	1.4	1.2	1.3	1.2

Table 2: Results of Experiment 2, Stage 4. Percentage of responses of different types.

Composite:	4-Morph		All Four		Best		Worst	
Array Type	Pres	Abs	Pres	Abs	Pres	Abs	Pres	Abs
Correct:	41	41	28	13	34	19	28	28
Incorrect:								
(1)Not picked	22		9		15		9	
(2)False choice	37	59	63	87	50	81	63	72

Footnotes

[i] Having chosen targets likely to be familiar to the majority of the participants, we did not make any formal check on their familiarity with the celebrities, since to do this at an early stage was difficult without the operator knowing which target they were viewing, and to do it at the end of the study risked wasting a substantive investment of operator and witness time if they were discarded at that stage.

[ii] As there were only four targets in total in this experiment we did not conduct a separate items analysis to examine F1 and F2 effects as we had for Experiment 1, but here took target as a factor within the overall subjects analysis – so that we could examine in this way whether effects were consistent across the different items within the experiment.