Parallel approaches to composite production: interfaces that behave contrary to expectation

Short title: Parallel composite systems

Charlie D. Frowd (1) – responsible for correspondence, etc. Phone: 01772 893439. Email: cfrowd@uclan.ac.uk
Vicki Bruce (2)
Hayley Ness (3)
Leslie Bowie (4)
Jenny Paterson (3)
Claire Thomson-Bogner (3)
Alexander McIntyre (3)
Peter J. B. Hancock (3)

(1) Department of Psychology
University of Central Lancashire,
PR1 2HE

(2) College of Humanities and Social Science
University of Edinburgh, EH8 9JU

 (3) Department of Psychology
University of Stirling, FK9 4LA

 (4) ABM UK
Stirling University Innovation Park, FK9 4NF

Abstract
This paper examines two facial composite systems that present multiple faces during construction to more closely resemble natural face processing. We evaluated a 'parallel' version of PRO-fit, which presents facial features in sets of six or twelve, and EvoFIT, a system in development, that contains a holistic face model and an evolutionary interface. The PRO-fit parallel interface turned out not to be quite as good as the 'serial' version as it appeared to interfere with holistic face processing. Composites from EvoFIT were named almost three times better than PRO-fit, but a benefit emerged under feature encoding, suggesting that recall has a greater role for EvoFIT than previously thought. In general, an advantage was found for feature encoding, replicating a previous finding in this area, and also for a novel 'holistic' interview.
 (131 words)

Witnesses and victims of serious crime may construct a visual likeness of a suspect's face. This is known as a facial composite and is typically obtained by describing the appearance of a suspect and selecting facial features: hair, face shape, eyes, nose, etc. Facial composites were originally the domain of artists, professionals who sketched with pencils or crayons, but other approaches were developed for those less artistic. Examples include Identikit and Photofit, available about 40 years ago. Research has identified problems with them, including both a limitation in the range of features (Davies 1983) and feature selection carried out in isolation from a whole face, a sub-optimal procedure: features are normally seen in the context of a whole face (Davies and Christie 1982, Tanaka and Sengco 1997). These issues appear resolved with the modern systems and very good likenesses are now possible (Cutler et al., 1988, Koehn and Fisher 1997, Davies et al., 2000).

E-FIT and PRO-fit are computerised versions of Photofit used by police forces throughout the world. They have been found to produce composites that are named about 18% of the time from laboratory witnesses working from a recent memory of a target face (Brace et al., 2000, Bruce et al., 2002, Davies et al., 2000, Frowd et al., 2004, 2005a), a finding which suggests that most composites go unnamed. The situation is more worrying, however, when a more realistic delay to construction is used. Research by Frowd et al. (2005b) found that less than 1% of composites from E-FIT and PRO-fit were correctly named with a 2 day delay. We note a similar finding for the Mac-A-Mug Pro, a sketch-based computerised system (Koehn and Fisher 1997).

Why might naming be so low for composites constructed after 2 days? Frowd et al. (2005b) proposed that this could be the result of a witness's memory becoming more of an impression after such a delay, with weakened access to facial features. Their work compared several systems including E-FIT, PRO-fit and a sketch artist. While performance was low after 2 days, composites from the sketch artist were better. Taken together with evidence that sketch is more of a holistic technique (Davies and Little 1990), a witness's memory of a suspect may be more holistic in nature after a relatively long interval. Frowd et al. also evaluated a novel system called EvoFIT, designed to be holistic in nature, and found that its composites were named better than those of E-FIT and PRO-fit (though not as good as those from a sketch artist).

EvoFIT does not require selection of individual features. Instead, there is a shape and pixel intensity model, built from whole faces using Principal Components Analysis, to provide a holistic coding system (e.g. Hancock et al. 1997). EvoFIT also capitalises on our relatively good ability to select faces that appear similar to a target (e.g. Hancock et al. 2000), a holistic operation. In use, witnesses peruse a range of faces and select a small number. These 'parent' faces are then bred together to combine their characteristics and produce another set. Repeating the selection and breeding process produces faces that more closely resemble the target face and, after three or four iterations, some very good likenesses can be evolved (e.g. Frowd et al., 2000).

EvoFIT also differs from the traditional approach by presenting more than one face at a time. Currently, 18 faces are seen together, the maximum that will sensibly fit on a computer monitor. It is possible that simply presenting more than one face at a time encourages holistic face processing and provides a better match with a witness's memory (an impression) after a couple of days. While this multi-face format may also encourage a relative judgement strategy, believed to cause false identifications in simultaneous police line-ups (Wells 1984), it may be valuable for composite construction where the task is to select the most similar exemplar. Note that this 'parallel' format is a U.K. requirement for witnesses inspecting a mugshot album for suspects.

We are aware of two earlier multi-face systems. Caldwell and Johnston (1991) presented sets of 20 faces with randomly selected features; users indicated the best and a composite was 'evolved' rather like EvoFIT. Rakover and Cahlon (1991) presented pairs of faces and built composites from features in faces selected most often. Unfortunately, neither system appears to have been the focus of a formal evaluation.

In the current study, we also evaluated a multi-face version of PRO-fit. Like E-FIT, PRO-fit normally uses a single face from which features are switched in and out. Our 'parallel' version presents an array of faces that differ in only one feature, for example hair, as seen in Figure 1. The interface thus allows the user to see and chose between multiple different versions of a feature. We note that this procedure is similar to that used with Photofit, where witnesses saw more than one feature at a time, but here they are seen in the context of a whole face..

The following five experiments evaluate EvoFIT and the parallel version of PRO-fit. Attention is given to PRO-fit in the first three experiments; EvoFIT in Experiments 4 and 5. In each experiment, it was predicted that a multi-face system would outperform a standard single-face one.

1. Experiment 1 – Parallel PRO-fit in a realistic setting

Experiment 1 compared the quality of composites constructed from both the standard (serial) and parallel PRO-fit user interface. Two phases were required for this evaluation. In the first part, participant-witnesses each constructed a composite with the help of a composite operator using either standard or parallel PRO-fit. In the second part, the composites were evaluated, initially by asking third persons to name them.

Figure 1. The PRO-fit facial composite system currently used to construct a facial composite (left) and a 'parallel' interface to the software that allows multiple examples of a feature to be presented simultaneously (right).
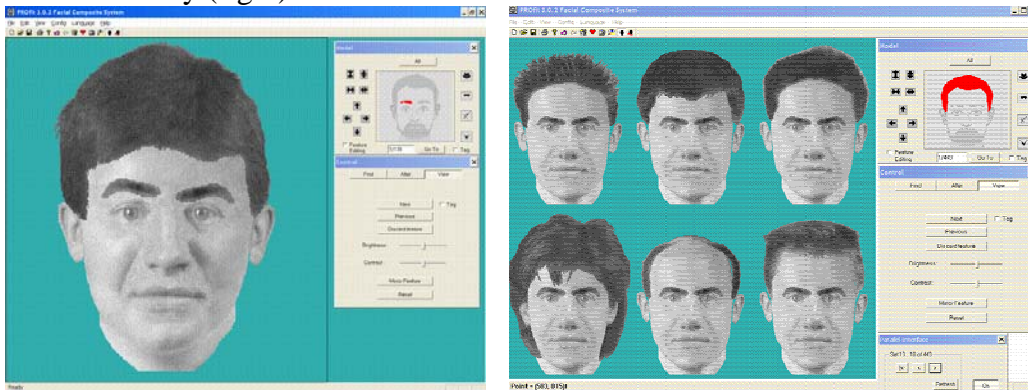
1.1. Composite construction

Research has shown that the person controlling the software, the composite operator, can affect the quality of a composite (Christie et al., 1981, Davies et al., 1983, Gibling and Bennett 1994). As this was a variable we were keen to limit, a single operator was used for each experiment but given appropriate training to permit consistent effects across experiments. In practice, each operator was trained 'in house', for all but Experiment 3, and then practiced extensively. Training was also given for the cognitive interview, a set of techniques to assist witnesses recall (e.g. Geiselman et al., 1986). The version of cognitive interview used in each experiment followed guidelines for UK criminal investigations (ACPO 2000) and included, *rapport building*, to facilitate recall; *reinstatement of context*, whereby witnesses form a mental image of the environment where the target was seen; *free* recall, the uninterrupted description of a target's face; and *cued* recall, whereby the operator prompts for information additional to free recall.

1.1.1. Target stimuli. Celebrity faces were chosen as targets to allow the composites to be named in the second part. However, participant-witnesses only built composites of faces they were unfamiliar with to more closely resemble the eyewitness scenario. Ten celebrity male photographs were located via an extensive search on the Internet and depicted each person (as far as possible) in a full-face pose and a neutral expression. The set were of young males ($M = 26$ years) to approximate suspect demographics (e.g. Goffredson and Polakowski 1995) and were well-known by our undergraduates who would carry out naming. Included were actors (David Boreanz, Scott Caan, Hayden Christiansen, Matt Le Blanc, Tobey Maguire and Elijah Wood), singers (Lance Bass, James Bradfield and Shane Filan) and a TV presenter (Anthony McPartlin).

Two sets of target stimuli were printed in colour on the same high quality printer (one set for each interface).

1.1.2. Participants. Twenty staff and students at Stirling University, 10 per interface, were paid £10. There were 14 females and six males from 18 to 52 years (*M* = 29.0 years, *SD* = 9.7).

1.1.3. Procedure. In brief, participant-witnesses inspected a photograph of famous male face for 1 minute, then 2 days later described his face via our cognitive interview and constructed a composite with the serial or parallel version of PRO-fit. This was achieved using two visits to the laboratory, first to inspect a target photograph, then to describe his face and construct a composite. Each person was tested individually.

Upon arrival at the first visit, participants were informed that famous faces were used as targets to allow the resulting composites to be named by third persons. However, as witnesses who construct composites are unfamiliar with suspects, participants would first locate a famous face that was unknown to them, and then study it for 1 minute. They were also asked not to reveal the identity of any famous face.

Participants were randomly assigned to construction by serial or parallel PRO-fit and were given one of two envelopes containing the relevant target photographs. The operator turned her back (so that the targets would not be seen) and participants were asked to select a photograph at random from the envelope. If the famous face was recognised, they were asked to return the photograph and select another. If all photographs were known, the participants were thanked and dismissed (three further participants were excluded in this way). Otherwise, they were given 1 minute to inspect the photograph. After reporting the target code, they then placed the photograph in a second 'used' envelope (i.e. non-replacement sampling was used).

Two days later, participant-witnesses returned to construct a composite. This was initiated by a 'rapport' building stage, to allow witnesses to relax, and involved the operator chatting informally for several minutes. Witnesses were informed that a two stage procedure would ensue, starting with a cognitive interview and followed by the construction of a composite.

The first phase was initiated by a short overview of the cognitive interview. Witnesses were then asked to think back to when the target had been seen and form a mental image of the room and his face. When witnesses stated that this had been achieved, they freely recalled details of his face. The operator explained that as more than one attempt at describing a face helps recall, they should repeat the exercise. Once complete, the session moved on to cued recall, with the operator reading back the description given for each feature and prompting for further details. When done, the session moved on to composite construction.

The operator gave an overview of PRO-fit plus a short demonstration of how features could be selected, resized and repositioned. It was explained that as the facial features were cut from photographs, only a general likeness may be possible, but a paint package was available to improve the quality. This utility allows part of a feature to be added or removed, useful for hair; additional lines to be added, for forehead wrinkles and under-eye bags; and the addition of general shading. It was noted that this artwork is normally carried out when all features have been selected to limit the need for re-work. Finally, as PRO-fit contains many examples of each facial feature, the witness's verbal description would be used to limit the number of features seen. Thus, the first stage involves the operator locating features to match the description. This would result in an 'initial' composite from which witnesses could suggest improvements.

The initial composite was thus prepared and presented to witnesses. They were given the freedom to decide which feature to work on, though this was normally hair and face shape initially, and selected the best from about two dozen examples presented. For those assigned to the parallel interface, examples of each feature were presented in sets of six faces; for the serial interface, examples were shown sequentially. Witnesses were given the opportunity to suggest changes to the size and position of each feature as well as to their brightness and contrast level. This process was repeated until witnesses decided that the best likeness had been achieved.

While participant-witnesses could request artistic enhancement throughout, as discussed above, they were given another opportunity at this stage. For the parallel PRO-fit group, this was carried out on a single face using standard PRO-fit. When complete, the resulting image was saved to disk as the composite.

1.1.4. Composite construction time. The time taken to construct a composite was faster for participant-witnesses using the parallel interface ($M = 37$ minutes) than the serial interface ($M = 43$ minutes), though this difference was not significant ($t_{18} = 1.34$; $p > 0.05$).

1.2. Composite naming

Participants naïve to the study attempted to name the composites. Two testing booklets were used, 10 composites in each, with each booklet containing five composites from the serial interface and five composites from the parallel interface.

1.2.1. Participants. Eighteen undergraduates at Stirling University named the composites. Their age ranged from 17 to 33 years with a mean of 25.4 years ($SD = 4.7$).

1.2.2. Procedure. Participants were tested individually, and randomly assigned, in equal numbers, to each testing booklet. They were told that they would be given a set of composites of famous faces and asked to provide the name of each famous person depicted. Participants were encouraged to guess when unsure.

Thus, composites from one booklet were presented sequentially for naming. No feedback was given as to the accuracy of the response. After all the composites had been inspected, naming was repeated for the target photographs. An *a priori* rule was applied such that only participants who knew at least five of the target photographs would be included in the study, as low target familiarity would limit composite naming (data from two further participants were excluded using this rule). The order of composites and targets was randomised after each person.

1.2.3. Results. Surprisingly, there were only five correct names elicited from the composites. These were distributed over four composites, with two correct names for composites from the serial interface, and three from the parallel interface. This low level of naming was in spite of high familiarity with the target photographs: participants correctly named them 72.8% of the time ($SD = 12.3\%$). Therefore, composites were correctly named only 3.8% of the time (i.e. discounting failures where the recogniser did not know the target and so could not name the composite). No inferential statistics were conducted due to the low values.

While the number of correct names was very low, participants produced an average of 3.6 incorrect names for the serial interface and 2.6 for the parallel interface; a significant difference using a within-subjects t-test ($t_{17} = 2.15$; $p < 0.05$).

1.3. Composite sorting

Given the very low level of correct naming, a composite sorting task was administered, requiring further participants to match the composites to their target photographs. Performance is much higher on this task and, despite its simplicity, serves as a good proxy to naming (Davies et al., 2000, Frowd et al., 2005a, 2005b). As participants inspected all composites, the design was within-subjects for interface type.

1.3.1. Participants. Eighteen staff and students at Stirling University volunteered. There were three males and 15 females and their age ranged from 18 to 38 years ($M = 27.2$ years).

1.3.2. Procedure. Participants were tested individually. They were told that they would be evaluating a set of facial composites constructed in a realistic study. The famous faces were introduced as the targets and placed on the table in front of them (in two rows of five). Next, a pile containing the 20 composites was given and they were asked to match them to the photographs. They were informed that there was more than one composite of each person and so should form a pile in front of each celebrity (though no mention was made regarding the number of repeats). They were also asked to sort each composite individually, in the order given, and try not to look at previous matches.

The order of the targets and composites were randomised after each person.

1.3.3. Results. The composites were sorted to mean accuracy of 38% for the parallel interface (*SD* = 13.4) and 51% for the serial interface (*SD* = 16.1); a significant benefit for the serial interface using a two-tailed within-subjects t-test (t_{17} = 3.51; $p < 0.005$).

1.4. Discussion

Participant-witnesses constructed a composite using the serial or parallel interface to PRO-fit. The resulting composites were correctly named only 3.8% of the time, a level too low to differentiate between the interfaces, though significantly fewer incorrect names were produced from the parallel interface. Composite sorting was used as an alternative and indicated that the serial interface was better.

 The low level of composite naming was unexpected, especially for the parallel interface, but does nonetheless reflect other research employing a 2 day delay (Koehn and Fisher 1997, Frowd et al., 2005b). It is curious that fewer incorrect names were produced for the parallel interface, which would normally be a good feature of a system as it avoids wasting police time, but is of little value here as correct naming was so low. However, composite sorting suggested that the parallel interface was inferior. In this task, participants tend to compare facial features between composites and targets and so it provides a general measure of feature quality (e.g. Frowd et al., 2005a). Thus, in addition to poor correct naming, the parallel interface prompted worse feature selection.

 A plausible explanation is that the parallel interface is harder to use than the serial interface after a 2 day delay, when a witness's memory is difficult to access for composite construction (Frowd et al., 2005b). If this were the case, then shortening the delay would reduce the task demands and improve performance. We note that should such an advantage exist, the parallel interface could function in criminal investigations where earlier construction was feasible.

 This notion was explored next for composites constructed with the target present, a procedure that should improve performance.

2. Experiment 2 – Parallel PRO-fit under ideal conditions

The second experiment did not involve witnesses; instead an operator constructed a set of composites using both interfaces with the target in view. This 'gold standard' procedure has been used elsewhere in this field (e.g. Cutler et al., 1988, Brace et al., 2000), and should elicit maximum performance. As a facsimile-like quality is possible with the target present using artistic paint tools, thus masking differences between interfaces, the operator did not undertake extensive artwork. Instead, elaboration was kept to a minimum, with most carried out for hair, as normal in real life. Two sets were produced in this way, one for each interface. For this study, a new operator was trained 'in house' and practiced until proficient.

2.1. Composite construction

The targets were the 10 celebrity faces used by Frowd et al. (2005b) and produced composites that were named less than 1% of the time for PRO-fit and E-FIT; their use here also provides an indication of likely system performance under more favourable conditions. The set was similar to Experiment 1, with a mean age of 26 years, and consisted of actors (Ben Affleck, Matt Damon, Jeremy Edwards, Joshua Jackson, Philip Olivier and James Redmond) and pop singers (Kian Egan, Mark Feehily, Ronan Keating and Ian 'H' Watkins). Two sets of photographs were printed (on the same high-quality printer) and placed in separate envelopes.

2.1.1. Procedure. The 10 targets were constructed twice, once for each interface. The operator selected a photograph at random, set the feature descriptors in PRO-fit and carried out construction using the serial or parallel interface (order randomised) as in Experiment 1. Artistic elaboration

was carried out as in Experiment 1, but was limited to minor adjustments mainly for the hair. Composites took about an hour to construct.

2.2. Composite naming

Composite quality was initially assessed by naming using two testing booklets and the procedure of Experiment 1.

2.2.1. Participants. Eight men and 16 women aged 18 to 44 years at the University of Stirling volunteered ($M = 22.0$ years, $SD = 5.7$). Data from six further participants were excluded as they failed to correctly name at least five target photographs.

2.2.2. Results. As expected, naming was much better at 20% correct. The average participant correct naming score (adjusted for recognition of the targets) was 20.4% ($SD = 24.9$) from the serial interface and 19.6% ($SD = 25.9$) from the parallel interface; a non-significant difference ($t_{23} = .11$; $p > 0.05$). The mean target naming was similar to Experiment 1 at 79.6% correct ($SD = 18.5\%$).

      The number of incorrect names per participant was much lower here but varied little over interface type: a mean of 0.9 ($SD = 1.1$) for the serial interface and 0.8 ($SD = 0.7$) for the parallel interface; a non significant difference ($t_{23} = 0.72$; $p > 0.05$).

2.3. Composite sorting

Composites were also assessed by sorting using the procedure of Experiment 1.

2.3.1. Participants. Seventeen staff and students aged between 20 and 67 years at the University of Stirling volunteered ($M = 36.2$ years, $SD = 19.5$). There were six females and 11 males.

2.3.2. Results. The composites were sorted to an accuracy of 72.1%. As before, more composites were correctly sorted from the serial interface ($M = 75.9\%$) than the parallel interface ($M = 68.2\%$); a significant difference ($t_{16} = 2.34$; $p < 0.05$).

2.4. Discussion

Composites in Experiment 2 used a method that should elicit maximum performance: a single operator attempted to copy features from photographs. As such, the procedure removes the difficulty of retrieving a face seen 2 days previously and should make construction easier. Accordingly, composite naming was found to be about 20%, clearly no longer at floor level, and emphasises the large effect of memory (refer to General Discussion). As in Experiment 1, naming did not differentiate the interfaces, but sorting favoured the serial interface. Thus, it would appear that the parallel interface was still inferior under a more optimal setting.

      It is still possible that the construction process was not ideal. The parallel interface was designed to be a better match with holistic face processing, but construction in this experiment encouraged a comparison of features between targets and composites: a feature task. A better paradigm might return to construction from memory, arguably more of a holistic task, but with a much shorter delay than 2 days.

      In the next experiment, witnesses constructed a composite immediately after seeing a target. We explored the notion that the parallel interface would be better if a witness's memory was more holistically biased. This was investigated by varying the method used to encode (learn) and then decode (recall) a target face.

3. Experiment 3 – encoding and interview

Laughery, Duval and Wogalter (1986) found that participants tend to remember a face by scrutinising the individual features if they know a face perception task will ensue. Such a feature-based method of encoding is likely to be of benefit when the task involves a serial recall exercise such as describing a face (Wells and Turtle 1988), or selecting features to construct a composite (Wells and Hryciw 1984), but not for recognition (e.g. Bower and Karlin 1974). In contrast, performance is known to be better in a recognition task when participants have encoded a face at a more global level: a holistic encoding (e.g. Shapiro and Penrod 1986).

These observations suggest that *encoding specificity*, proposed by Tulving and Thomson (1973), may be relevant here, as retrieval is likely to be better if the conditions at encoding (target exposure) match those at decoding (composite construction). Given that participants in Experiment 1 were likely to have carried out a feature encoding, as suggested above by Laughery et al. (1986), and that the operator compared features in Experiment 2, it is not surprising that a benefit emerged for the serial interface if we were correct to assume that it is a more feature-based approach. Thus, feature encoding may be best for serial PRO-fit but holistic encoding may be best for parallel PRO-fit.

The manipulation of target encoding used here was similar to Wells and Hryciw (1984). Their participants viewed a target face under feature encoding, by rating on 10 physical dimensions, or trait encoding, by assigning 10 personality traits. Participants then either constructed an Identikit composite or attempted to recognize the target from a lineup. They discovered that construction was best under feature encoding but identification was best under trait encoding. The benefit of feature encoding on composite production has since been replicated (Laughery et al., 1986).

In spite of a change in encoding, the parallel interface might still suffer interference, this time from witnesses describing their target face. As mentioned above, verbal description production is another feature-based task and has been found to bias the cognitive system to the detriment of face recognition (e.g. Schooler and Engstler-Schooler 1990, Westerman and Larsen 1997). Therefore, it is prudent to change the type of interview given to witnesses using the parallel interface. In this case, an interview that encouraged holistic processing would appear appropriate.

Thus, three factors were manipulated: target encoding (feature / holistic), interview (cognitive / holistic) and interface (serial / parallel). The parallel interface was expected to be better than the serial interface under holistic encoding and a holistic interview, but worse under feature encoding and a cognitive interview (and vice versa).

3.1. Method

In the current work, participants either assigned a single personality judgement (holistic encoding) or inspected the features of a target face (feature encoding) while watching a video. They then underwent either a cognitive interview of the type used previously, or a 'holistic' interview that required them to make a series of personality judgements. The final stage was the construction of a composite.

The interviews were designed to be as similar to each other as possible. Thus, witnesses receiving a holistic interview first described in their own words the personality of their target face and were then prompted for a number of personality judgements. As the cognitive interview elicits information on seven facial features (hair, face shape, brows, eyes, nose, mouth and ears), the holistic version requested seven personality traits (honesty, intelligence, friendliness, kindness, excitability, selfish and arrogance). To further encourage holistic processing, those receiving a holistic interview also assigned a rating for each trait.

The new interview required a new procedure. Recall that construction normally begins with an operator preparing an 'initial' composite, a face with features that match a witness's description. However, a feature-based description was not produced for a holistic interview, and so construction began from the so-called 'default' composite, the face presented when PRO-fit starts – it contains features listed first in the database. For these witnesses, the operator merely presented

features in the order listed in the database and then located more specific examples based on feedback.

The experiment employed a third composite operator. She attended a recognised training course for UK operators and is experienced, having constructed composites for other research projects (e.g. Bruce et al., 2002, Frowd et al., 2005a, 2005b).

3.2. Target stimuli

For convenience, four video clips were used, each lasted 30 seconds and depicted a different male in a head-and-shoulders pose looking briefly to their left, right and finally speaking directly into the camera for about 15 seconds (sound was muted for participants). The sequences involved an unfamiliar male and so an additional target familiarity check was not necessary.

3.3. Composite construction

The design manipulated witness encoding (feature / holistic), interview (cognitive / holistic) and interface (serial / parallel), and each target face was constructed once in each condition. An additional development was made, to encourage holistic processing further, by doubling the number of faces presented in the parallel interface to 12 faces.

3.3.1. Participants. Thirty-two staff and students at the University of Stirling were paid £10. There were 13 males and 19 females and all were less than 40 years of age.

3.3.2. Procedure. Participants were randomly allocated to encoding (feature / holistic), interview (cognitive / holistic) and interface (serial / parallel) conditions: four composites were constructed (one per target) in each of these eight conditions (2 feature x 2 interview x 2 interface); a total of 32 composites. Recruitment to the study was via a small advert for a face perception experiment (i.e. they did not know that a composite was required).

Each person was tested individually. Upon arrival, they were invited to watch a short video. Those in the feature encoding group were asked to spend a few seconds studying each facial feature of the person depicted in the video; for holistic encoding group, participants were asked not to study his facial features, but to watch the whole clip and then rate the perceived generosity of the person (1 = *not generous* / 10 = *extremely generous*). Afterwards, participants were informed that they would be constructing a composite of this person. It was explained that one of the objectives of the study was to compare witnesses who remember their target face by features or by a holistic judgement; hence the need for the prior instructions.

Participant-witnesses receiving a cognitive interview were given the procedure of Experiment 1 (free recall then cued recall). A similar format was followed for the holistic interview group: they were informed that they would be required to describe the personality of target. Initially, this would be done as a 'free' description but, thereafter, they would be asked to make a number of personality judgements (1 = *negative attribute* / 10 = *positive attribute*).

Witnesses were asked to form a mental image of his face, then describe his personality (in general, they found this free recall exercise very difficult). After providing a description, they gave a rating for honesty. They were then asked to form a mental image of the target face again, and the operator moved onto the next trait. The operator worked sequentially through the personality traits, prompting witnesses to give a description and a rating. The other traits were intelligence, friendliness, kindness, excitability, selfishness, and arrogance.

The session moved on to composite construction. For participants receiving a cognitive interview, construction followed the procedure of Experiment 1, except that parallel PRO-fit displayed 12 faces at a time. Construction was also the same for those in the holistic interview group, except they were told that the first face shown would be the 'default' composite face, the image presented when PRO-fit is started, and about 30 alternatives of each feature would be seen as listed in the database. Then, more specific features would be presented based on their feedback.

Thus, the operator displayed the default face and a composite was constructed using this procedure.

3.4. Composite sorting. The composites were of unfamiliar faces and so a sorting task was used for evaluation. A good quality monochrome full-face image still was taken from each of the four videos and used as the target photographs for this task. The same procedure as Experiment 1 was followed, except that participants were told that the composites were of unfamiliar faces.

3.4.1. Participants. Twelve staff and students at Stirling University volunteered. There were four males and eight females from 22 to 41 years. The mean age was 27.8 years ($SD = 5.0$).

3.4.2. Results. Overall, participants sorted the composites to an accuracy of 58.1%. Figure 2 shows performance by interview and interface; the effect of witness encoding is not shown for simplicity as it was found to be consistent (see below). Specifically, composites from the serial interface ($M = 62.5\%$, $SD = 22.5$) were sorted better than from the parallel interface ($M = 53.6\%$, $SD = 24.7$); composites following a feature encoding ($M = 63.5\%$, $SD = 24.1$) were better than following a holistic encoding ($M = 52.6\%$, $SD = 22.6$); and there was little difference between the two types of interview (cognitive interview: $M = 59.4\%$, $SD = 24.5$; holistic interview: $M = 56.7\%$, $SD = 22.6$).

Figure 2. The effect of interview and PRO-fit interface on composite quality (as assessed by a composite sorting task).
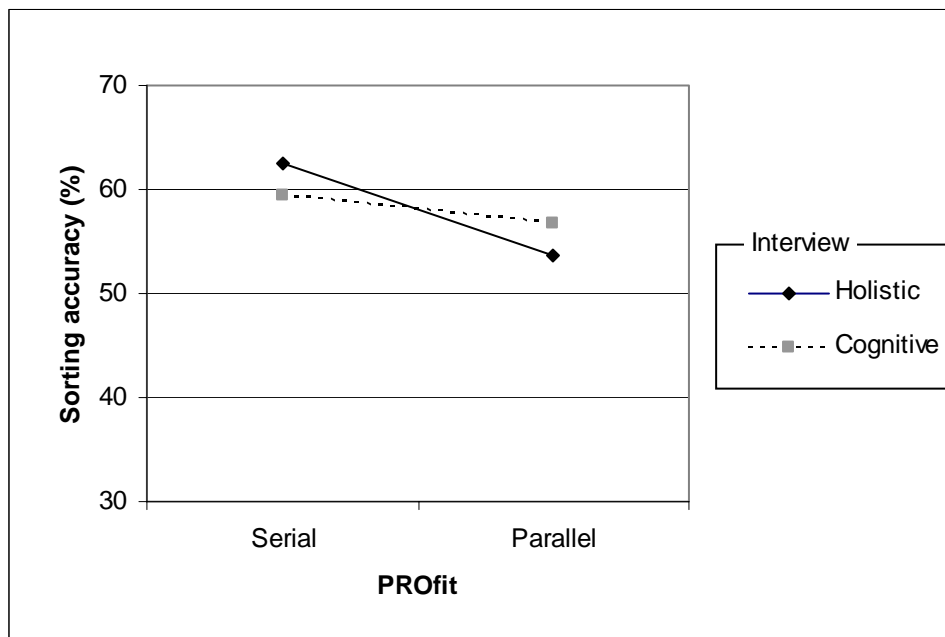
   Note. For simplicity, the data has been collapsed over witness encoding (since feature encoding was found to be consistently better than holistic encoding).

The participant sorting accuracy scores were subjected to a three way repeated-measures ANOVA. This was significant for interface type ($F_{1,11} = 6.5$; $p < 0.05$) and encoding ($F_{1,11} = 11.2$; $p < 0.01$) but interview was not significant ($F_{1,11} = 0.32$; $p > 0.05$). There was only one significant interaction, between interface and interview ($F_{1,11} = 7.8$; $p < 0.05$), as shown in Figure 2; both interface and encoding ($F_{1,11} = 1.5$; $p > 0.05$), and interface, interview and encoding ($F_{1,11} = 0.1$; $p > 0.05$) were not significant. Simple-main effects of the significant interaction revealed that composites following a holistic interview were better from the serial interface than the parallel interface; composites from the parallel interface were better following a cognitive interview (cf. a holistic interview); and, although composites from the serial interface were better with a holistic interview than a cognitive interview, this just fell short of significance ($p = 0.06$). Note that there

10

was no significant difference between the interfaces following a cognitive interview (dotted line, Figure 2).

3.5. Discussion

Experiment 3 manipulated target encoding, interview and interface. Participants watched a short video of an unfamiliar male in one of two encoding conditions (feature / holistic), then received a holistic or cognitive interview and constructed a composite using serial or parallel PRO-fit. The sorting task indicated that whereas the parallel interface was similar to the serial interface under a cognitive interview, it was worse under a holistic interview. Sorting also found that feature encoding was better than holistic encoding; and this was unexpectedly consistent across interviews and interfaces.

Once again, the parallel interface was found not to be more of a holistic system than the serial version: no advantage emerged for holistic encoding, and performance was worse following a holistic interview; in fact, under the strongest holistic influence, a holistic encoding and a holistic interview, the parallel interface performed the worst.

The first three experiments indicate that the parallel interface does not behave as expected and is not quite as good as the serial interface: it was worse following a realistic 2 day delay (Experiment 1), target present construction (Experiment 2), a holistic encoding and a holistic interview (Experiment 3). In general, the interface appears to encourage a poorer selection of features, based on worse sorting scores, and suggests that merely presenting multiple examples of a facial feature, albeit in the context of a whole face, is not beneficial. While the following work now considers the other parallel system, the General Discussion argues that the heart of the problem with parallel PRO-fit may be the high degree of similarity between the presented faces.

EvoFIT faces do not change by a single feature; instead, the whole face changes under the global influence of the underlying parameters. While there are other differences between EvoFIT and parallel PRO-fit, such as an underlying breeding process, the increase in variability between presented faces may be important. Indeed, previous work has shown that EvoFIT can produce more identifiable composites than E-FIT and PRO-fit, albeit with low naming levels (Frowd et al., 2005b). Since then, several key enhancements have been made (see below) and the following work assesses current performance. The next experiment follows on from the previous, using a similar design and target set; Experiment 5 involves a more realistic study with a 2 day delay and composite naming. In both cases, performance is compared against PRO-fit.

4. Experiment 4 – EvoFIT under different encoding

While it is not within the scope of this article to discuss all the technical aspects of EvoFIT, which may be found in Frowd et al. (2004), the key principles and recent enhancements will be outlined here. EvoFIT contains a pair of face models: the first, referred to as 'shape', models the shape of facial features and the distances between them; the other, known as 'texture', is a model of grey scale pixel values. To generate a face with random characteristics, a random texture is first generated from the texture model and then morphed to a random shape from the shape model. A set of about 70 faces are initially generated in this way, each with a hairstyle selected from PRO-fit. An example screen, showing a typical set of 18 faces, may be seen in Figure 3.

Figure 3. The EvoFIT facial composite system: witnesses repeatedly select from about 70 faces (the first 18 are shown here) that are bred together to improve the likeness.
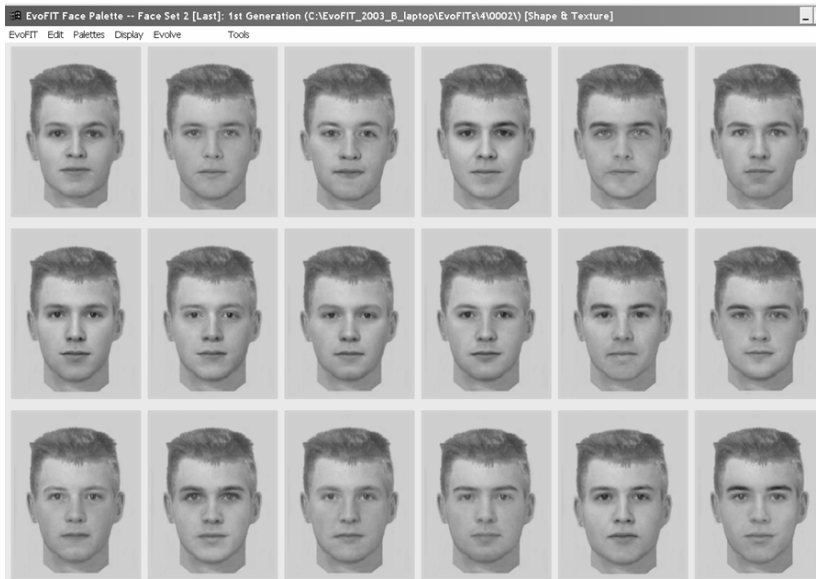
Composite construction with EvoFIT is perhaps best described as a search problem: there exists a face with a good likeness within the system, the problem is locating it. Evolutionary Algorithms are primarily used for this purpose (e.g. Mitchell 1996), involving the selection and breeding of faces, though several mechanisms serve to accelerate the process. For example, witnesses first select facial shapes then facial textures. They also nominate the closest shape and texture, to form a 'best' face, which is both given more breeding opportunities and propagated unchanged into the next generation. In addition, to assist with selection, facial textures are presented with a previously selected best shape, and vice versa (except in the first generation, as shapes are selected first). Further, although witnesses are requested to focus on the overall appearance of a face, they sometime request specific changes such as widening a face or moving the eyes together. A feature 'shift' tool was thus designed to allow such changes to the best face.

EvoFIT has undergone a number of formal evaluations (Frowd et al., 2000, 2004, 2005a, 2005b), the most recent of which suggests that EvoFIT produces slightly more identifiable composites than the other UK computerised systems. Since then, the EvoFIT face model has been enhanced to reduce occasional image distortions and improve image quality. Also, given the relative importance of the best face for evolution, as described above, the identification of this face has been improved by presenting all combinations of selected shapes and textures (previously, it was difficult to see which shape and texture might be combined to give the best overall likeness).

To evaluate the new version of software, an experiment similar to the previous was used. EvoFIT was tested under two conditions designed to produce 'extremes' of performance: a holistic encoding and a holistic interview, predicted to be the best, and a feature encoding and a cognitive interview, predicted to be the worst. These composites were compared against another set using PRO-fit. This time, the serial version of PRO-fit was used with a cognitive interview and feature encoding; and parallel PRO-fit was used with a holistic interview and holistic encoding. This allowed EvoFIT to be evaluated with PRO-fit in 'normal' use (serial PRO-fit / feature encoding / cognitive interview), but would verify any benefit against the 'worst' PRO-fit scenario as suggested in Experiment 3 (parallel PRO-fit / holistic encoding / holistic interview). A fourth composite operator was trained 'in house' and was suitably experienced.

4.1. Composite construction

Construction was a 2 (encoding) x 2 (system) between-subjects design. The same stimuli as Experiment 3 were used, four 30s videos, but this time, each target was constructed eight times, twice per condition, to produce a total of 32 composites.

4.1.1. Participants. Thirty-two staff and students at Stirling University were paid £5. There were 11 males and 21 females, aged from 17 to 57 years ($M = 31.1$ years, $SD = 10.2$).

4.1.2. Procedure. Participants were randomly assigned to encoding (feature / holistic) and system (EvoFIT / PRO-fit) conditions. The procedure of Experiment 3 was followed for PRO-fit witnesses, except that a cognitive interview was used for feature encoding, and a holistic interview was used for holistic encoding. In addition, serial PRO-fit was used for feature encoding and parallel PRO-fit was used for holistic encoding. Note also that the eight targets were repeated twice in each condition to produce 32 composites in total (2 encoding x 2 system x 4 targets x 2 repeats).

Participant-witnesses assigned to EvoFIT followed the same procedure as the PRO-group for target exposure and interview. They were then given an overview of the construction procedure, emphasising that hair would be selected first using PRO-fit, then a composite would be evolved using EvoFIT.

An overview of PRO-fit was given, relevant to hair selection. For participants in the cognitive interview group, about two dozen hairstyles were presented to match the witness's description. For the holistic interview group, the first 30 or so hairstyles were shown, to identify a style, followed by about two dozen examples in that style. The operator encouraged witnesses to select the best hairstyle and used the artistic paint tools where necessary. When complete, the hairstyle was imported into EvoFIT.

The operator introduced EvoFIT and emphasised the basic process of selection and breeding. Witnesses were told that facial shapes would be selected first, then facial textures. For each facial type, they would inspect about 70 examples, over four screens of 18 faces, and select six faces. In addition, faces would have random characteristics initially but would resemble their target over time, with further selection and breeding.

The first screen of facial shapes was displayed. Witnesses were told that as textures had yet to be selected, these faces were given an 'average' texture. They were asked to select the best two shapes from each presented screen. Each time, they should base selections on the overall similarity with their target. Also, selected faces could be de-selected later. Thus, three screens of random shapes were presented and witnesses selected as requested. Another set was shown, but this time exchanges were made to ensure only six faces were selected. Finally, the best shape was nominated, and used to present the facial textures in the next part. Next, the facial textures were presented over four screens and witnesses selected as for shape. They were then shown 36 'combination' faces: all possible pairings of their six selected shapes and six selected textures. These were shown over two screens and witnesses identified the overall 'best' face.

The selected faces were bred together to produce offspring faces. The selection procedure was as before – for shapes, textures and combination faces – except that shapes were given the texture of the 'best' face, to assist selection. Also, witnesses considered whether a better shape had been evolved, in the presence of the best face. If so, this shape was used for the evolved textures; otherwise the shape of the best face was used.

At this stage, an opportunity was given to improve the likeness of the best face. If changes were required to the shape of features, or the relationship between them, the best face was taken into a small utility called the Feature Shift tool. The operator used this tool to adjust the size and spacing of features as requested.

The selected faces were bred together again, with face selection and Feature Shift use as before. This process was repeated until witnesses decided that the best likeness had been achieved. Finally, an opportunity was given to artistically enhance their composite, the same as for PRO-fit witnesses, but using Adobe Photoshop.

4.2. Composite evaluation

The design and procedure of Experiment 3 was used to evaluate the composites with a sorting task.

4.2.1. Participants. Twelve Stirling University students were paid £2. There were seven females and five males, aged 18 to 27 years ($M = 22.6$ years, $SD = 3.1$).

4.2.2. Results. Curiously, there was no difference in composite quality between systems ($M = 64.0\%$ for both), though there was a large benefit for feature encoding ($M = 75.0\%$ vs. 53.1%). Accordingly, a repeated-measures ANOVA was not significant for system ($F_{1,11} = 0.00$; $p > 0.05$), but indicated that feature encoding was best ($F_{1,11} = 46.2$; $p < 0.001$); the interaction between these factors was not significant ($F_{1,11} = .21$; $p > 0.05$).

4.3. Discussion

In this experiment, participant-witnesses constructed a composite with EvoFIT or PRO-fit following either a holistic encoding and a holistic interview, or a feature encoding and a cognitive interview. Sorting indicated that feature encoding was best for PRO-fit, supporting the above finding, but curiously, it was also best for EvoFIT. Recall that EvoFIT was expected to be best under a holistic encoding and a holistic interview, though was as bad as parallel PRO-fit.

Frowd et al. (2005b) found that EvoFIT was superior to PRO-fit and E-FIT by naming, but not by sorting. Thus, it is possible that the instrument used here for evaluation, a feature-based task, was not appropriate for EvoFIT. In the following experiment, composite naming was used, also along with a 2 day delay to provide a better match between a witness's memory and the EvoFIT system.

5. Experiment 5 – EvoFIT in a realistic setting

This experiment manipulated target encoding, as before, but included a 2 day delay and composite naming. The design featured a cognitive interview throughout, and serial PRO-fit under both encoding conditions, to evaluate the version of PRO-fit used by police. In Experiment 3, there was a sizeable advantage for feature encoding using a sorting task and one would expect this benefit to extend to composite naming.

5.1. Composite construction

Construction was a 2 (encoding) x 2 (system) between-subjects design. The realistic methodology of Experiment 1 was employed, including a 2 day delay.

5.1.1. Target stimuli. Well-known white male UK footballers were used. This assisted recruitment by allowing mostly female participant-witnesses, given their low familiarity with footballers and the finding that witness gender does not appear to affect composite quality (Frowd et al., 2005a, 2005b). Twelve photographs of footballers were selected via an extensive Internet search (as far as possible) in a full-face pose and a neutral expression. These were Tony Adams, Peter Beardsley, Dennis Bergkamp, Ole Gunner Solskjaer, Roy Keane, Steve McManaman, Emmanuel Petit, Wayne Rooney, Paul Scholes, Teddy Sheringham, Alan Smith and Zenadine Zidane. Four sets were printed using the same high quality colour printer (two sets each for PRO-fit and EvoFIT).

As static photographs were used in conjunction with a holistic encoding, participants made multiple personality ratings during target exposure, similar to Wells and Hryciw (1984), rather than a single rating as in Experiment 3 and 4.

5.1.2. Participants. Forty-eight staff and student at the University of Stirling were paid £10. There were 33 females and 15 males from 19 to 56 years ($M = 32.0$ years, $SD = 12.5$).

5.1.3. Procedure. Participants were informed that the study used football players, to enable the composites to be evaluated by naming. They were also made aware that exclusion would occur if

all targets were recognised as real witnesses who make composites do not know the identity of the suspect (seven further participants were excluded on this basis).

Each person was randomly assigned to encoding (feature / holistic) and system (PRO-fit / EvoFIT) conditions. The presentation of target photographs was carried out by another experimenter, as the procedure was deemed too complex for the operator, who must remain blind to the targets; otherwise, this part was the same as Experiment 1 until an unfamiliar target was located. Participants in the feature encoding group were given 1 minute to look at the photograph and study his facial features. For those in the holistic group, the first unknown footballer was immediately placed face down on the table. They were told to make 11 personality traits in the presence of the face, plus one other rating, each on a seven point scale (1 = *low* / 7 = *high*). The 12 judgements were read aloud: honesty, intelligence, aggressiveness, kindness, generosity, imaginativeness, arrogance, friendliness, selfishness, excitability, dislike and distinctiveness. They were instructed to look continuously at the photograph and provide a rating every 5 seconds. Thus, the photograph was turned over and the experimenter ensured that ratings were given in 5 second intervals and that the target exposure was 1 minute.

Witnesses returned to the laboratory 2 days later and met with the composite operator. The procedure followed Experiment 1 for those assigned to serial PRO-fit: they received a cognitive interview and constructed a composite using serial PRO-fit. The other group were given a cognitive interview as in Experiment 1 and then followed the procedure for constructing an EvoFIT detailed in Experiment 4.

5.2. Composite naming

Composites were initially assessed by naming using the procedure of Experiment 1, except that football fans were recruited. Also, to increase statistical power, a within-subjects design was employed: participants inspected all 48 composites.

5.2.1. Participants. Twelve football fans at Stirling University volunteered. There were eight males and four females aged from 17 to 25 years ($M = 19.8$ years, $SD = 2.4$).

5.2.2. Results. Roughly a third of the composites (31%) were correctly named by at least one person, distributed slightly in favour of EvoFIT (9/24 vs. 6/24). As can be seen Figure 4, the mean participant percent correct scores were higher for EvoFIT than PRO-fit ($M = 8.5\%$ vs. 3.7%), and were higher for feature than holistic encoding ($M = 8.0\%$ vs. 4.3%). A repeated-measures ANOVA confirmed these observations, with EvoFIT better than PRO-fit ($F_{1,11} = 8.6$; $p < 0.05$) and feature encoding better than holistic encoding ($F_{1,11} = 4.2$; $p < 0.01$); the interaction was not significant ($F_{1,11} = 2.3$; $p > 0.05$). The mean correct target naming was 93.8% ($SD = 10.7$).

The mean incorrect names per participant varied little by system (EvoFIT = 47.8%, PRO-fit = 49.5%) or encoding (Feature = 46.1%, Holistic = 50.9%). Indeed, an ANOVA found no main effect of system ($F_{1,11} = 0.09$; $p > 0.05$) nor encoding ($F_{1,11} = 0.88$; $p > 0.05$); the interaction was also not significant ($F_{1,11} = 0.23$; $p > 0.05$).

Figure 4. Naming of EvoFIT and (serial) PRO-fit composites from witnesses under feature and holistic encoding.
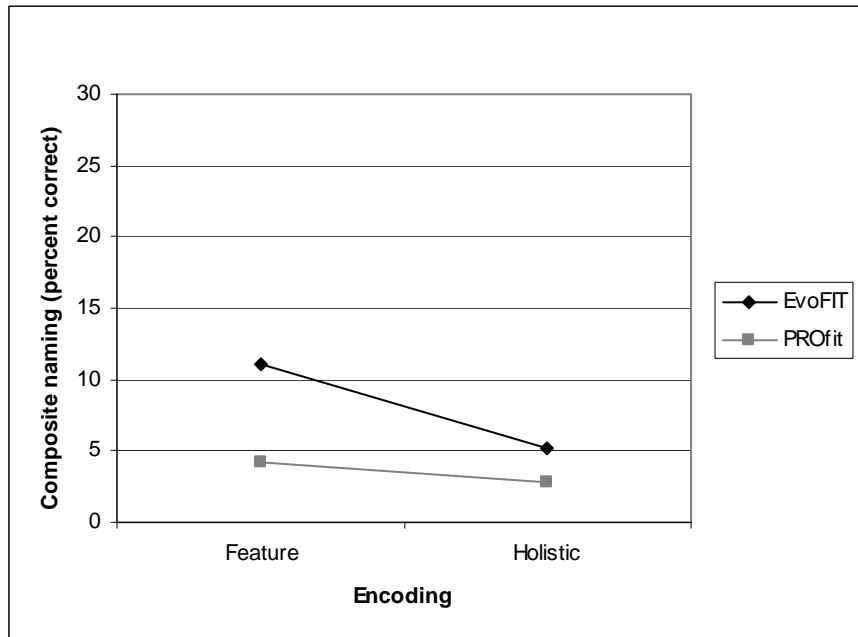
5.3. Composite sorting

Composites were also assessed by sorting using the procedure of Experiment 1.

5.3.1. Participants. Thirty-three undergraduate students from Stirling University volunteered to sort the composites.

5.3.2. Results. Composites were sorted best when made after a feature encoding ($M = 47.2\%$ vs. 42.2%) but, unlike naming, there was little difference between PRO-fit and EvoFIT ($M = 45.5\%$ vs. 44.7%). An ANOVA again confirmed a feature encoding advantage ($F_{1,58} = 106.3$; $p < 0.001$) and also that the systems were very similar ($F_{1,58} = 0.0$; $p > 0.05$). However, there was a significant interaction this time ($F_{1,58} = 13.7$; $p < 0.001$), as can be seen in Figure 5, as PRO-fit was better than EvoFIT under feature encoding ($p < 0.001$) but EvoFIT was better than PRO-fit under holistic encoding ($p < 0.05$); both systems exhibited a feature encoding advantage ($p < 0.005$).

5.4. Discussion

Experiment 5 manipulated target encoding and system in a design with a 2 day delay. As expected, naming was best for EvoFIT and reflects previous work (Frowd et al., 2005b). Also, the standard version of PRO-fit was best under feature encoding for composite naming and sorting, a result that replicates Experiment 3 and other research involving Identikit (Wells and Hryciw 1984, Laughery et al., 1986). Curiously, EvoFIT was also best under feature encoding, with naming more than double that of holistic encoding ($M = 11.8$ vs. 5.3%). However, while sorting showed an increase for EvoFIT under feature encoding (5%), it was significantly larger for PRO-fit (18%), and appropriately suggests that EvoFIT is less of a feature-based system than PRO-fit.

Figure 5. The effect of encoding on EvoFIT and (serial) PRO-fit composites (as measured by a sorting task).
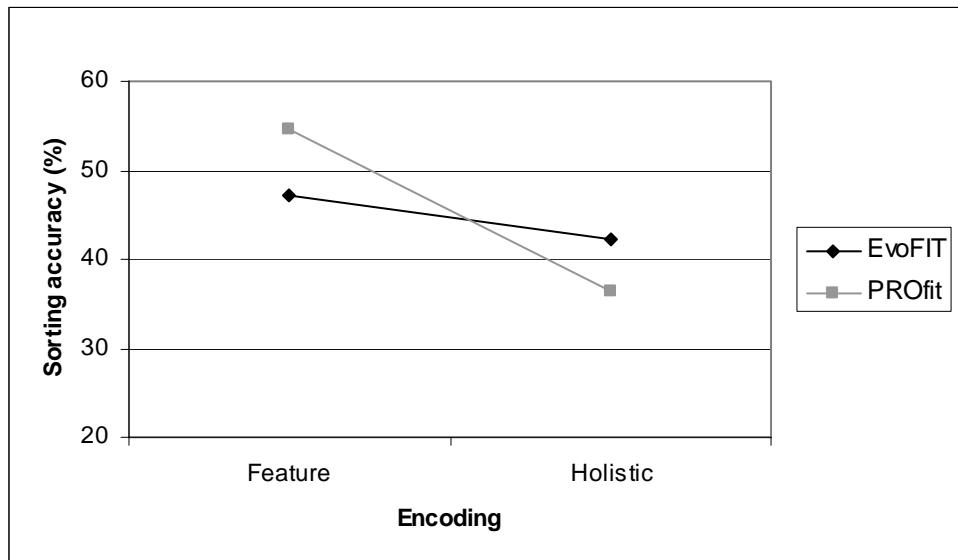
6. General discussion

This paper evaluated two multi-face composite systems. The first was a version of PRO-fit that displays features in sets of (initially) six. In Experiment 1, composites from the standard and 'parallel' PRO-fit turned out to be poorly named, but a sorting task favoured the standard system. Experiment 2 involved construction under more favourable conditions, to give better naming, but the same conclusion was reached. Experiment 3 manipulated target encoding and type of interview, likely sources of interference, but found that the new interface was worse under a holistic influence.

The next experiments considered EvoFIT. Unexpectedly, Experiment 4 found both a feature encoding advantage for EvoFIT and equivalent performance with PRO-fit. In Experiment 5, under a 2 day delay, EvoFITs were named almost three times better than those from standard PRO-fit, though a feature encoding advantage still remained for EvoFIT.

In summary, while the parallel interface to PRO-fit was not quite as good as the serial version, EvoFIT was better than PRO-fit by composite naming. Overall, composite naming was low following a 2 day delay and all systems tested were best under feature encoding, though EvoFIT demonstrated less benefits of feature coding, consistent with it being a more holistic system.

6.1. The parallel interface to PRO-fit

Our motivation for parallel PRO-fit emerged from composite systems that present more than one face at a time. The new system presented multiple variants of a single feature and, as such, is rather similar to the old Photofit system but using a whole face context. However, the interface was found not to be closer to natural face processing, as its composites were better neither after a holistic encoding (e.g. Wells and Hryciw 1984, Shapiro and Penrod 1986), nor after our holistic interview. We note that it did not even enhance feature processing either, given that there was no difference between interfaces following a cognitive interview.

As the procedure for testing the two versions of PRO-fit was the same, our data suggests that the very problem with the parallel interface is the parallel presentation of features! This is perhaps best explained by interference to the facial context. It would appear that presenting faces that change by a single feature produces a very similar set of faces - as illustrated in Figure 1. This high degree of similarity may have encouraged witnesses to focus on the individual feature being manipulated, thereby engaging in a more feature-based method of processing. Such an approach

would weaken the benefit of a whole face context and promote an isolated feature format, used previously with Photofit and Identikit and known to be sub-optimal (Davies and Christie 1982, Tanaka and Farah 1993, Tanaka and Sengco 1997).

This theory explains why sorting was consistently worse for parallel PRO-fit (the chosen features were worse because of a poorer context for selection) and why the holistic interview did not improve performance (interference to the whole face context encouraged a feature based approach). Such a theory is testable, comparing construction using the parallel interface with isolated feature selection, where equivalent results are expected. Such results may also be of value to a police line-up, or mugshot albums, suggesting that a sequential presentation may be preferable to a simultaneous one, especially when the presented faces are very similar to each other.

The interface may have been better if the presented faces varied more, perhaps by changing more than one feature at a time. However, merely presenting all feature combinations that match a witness's description is impractical: there would be too many faces. An alternative might be a Genetic Algorithm (GA), as used by Caldwell and Johnston (1991), with features sampled randomly from a composite system and then more faces bred based on user feedback. This is the general EvoFIT approach – large variability between faces, a parallel interface and a GA – found to be better here.

So, might there be a role for parallel PRO-fit in police work? While our results indicate poorer performance, differences were quite small (sorting: 13% in Experiment 1, 8% in Experiment 2, and 9% in Experiment 3). It is perhaps worth mentioning that police operators have expressed an interest in the new interface for exploring PRO-fit's feature library and are considering using it to assemble the 'initial' composite.

6.2. EvoFIT

The data suggests that in general EvoFIT does not contain superior features than PRO-fit, as sorting was similar, but it does appear to be a better match with a witness's memory, as naming was better. However, EvoFITs were named more than twice as often as PRO-fits under feature encoding and this suggests that EvoFIT retains a significant recall component, since previous work has suggested that feature encoding is better for feature selection (Wells and Hryciw 1984).

EvoFIT does involve an element of recall: the Feature Shift tool used during construction to enhance a best face. It was introduced early in development in response to user feedback and tests indicated that it was well received. In general, the tool facilitates a search of 'face space' by gravitating solutions towards a better likeness. Of course, witnesses must use language for its use and this recall process is likely to be enhanced following a feature-based encoding (e.g. Wells and Turtle 1988).

In spite of this unexpectedly strong recall component, EvoFITs were named better than PRO-fits after a 2 day delay. Although caution must be used when comparing across studies, it would appear that the enhancements made since the last evaluation (Frowd et al., 2005b) have improved naming. Of the two improvements made, the enhanced selection of the best face, by taking combinations of shape and texture, appears to be the most important; in previous work, the other change, to improve the face model, despite producing clearly better face images, did not promote a better composite (Hancock and Bruce 2003).

6.3. A problem of memory

A much better naming level of 20% was found for composites constructed by the operator in Experiment 2. While this procedure was not intended to reflect construction with real witnesses, other research with short delays has reported similar naming using a realistic paradigm: experienced operators, a cognitive interview, unlimited time for construction and artistic enhancement (e.g. Bruce et al., 2002, Frowd et al., 2005a). These data suggest that top-end performance is likely to have been reached, and at a level that is far from ideal. Thus, while current systems are not able to produce good renditions, there is also a limitation in a witness's memory.

The poor naming for PRO-fit in Experiments 1 and 5, with the relatively better naming for EvoFIT, supports the theory by Frowd et al. (2005b) that a witness' memory is more of an impression after a couple of days and holistic systems such as EvoFIT (and a sketch artist) will produce better renditions. After longer intervals, there appears to be limited access to individual features and a decline in performance for the feature-based systems, such as E-FIT and PRO-fit, relative to the holistic systems. This notion fits well with the observation that face recognition ability remains stable for a week or so (e.g. Laughery et al., 1974, Shepherd 1983), but recall, such as describing a face, falls off far more rapidly (e.g. Ellis, Shepherd and Davies 1980).

6.4. New procedures

The current work provided evidence that a current composite system, PRO-fit, has a substantial recall component, as found for Identikit (e.g. Wells and Hryciw 1984). In general, PRO-fit performed better under a feature encoding; it was also better than EvoFIT in a sorting task (Experiment 5). However, Experiment 3 found benefit for our holistic interview, with PRO-fit witnesses describing and rating the personality of their target and constructing a composite from the 'default' face. It has recently come to our attention that similar benefit of trait attribution has been found elsewhere (McQuiston 2003), suggesting that the effect may be reliable.

It is possible that the absence of a feature-based description for the holistic interview may avoid a *verbal overshadowing effect* (VOE), occurring when face description interferes with face recognition (e.g. Schooler and Engstler-Schooler 1990). While a VOE is undesirable, it is unlikely to affect composite quality given its small effect size (e.g. Meissner and Brigham 2001); it can also be rather elusive (Clifford 2002). So, if trait attribution is responsible, then it may be better to enhance a witness's recognition ability, rather than their recall – perhaps a valuable pursuit if memory is indeed more holistic in nature after a couple of days. We note that were this procedure to be used with 'real' witnesses, a different set of personality traits would be required, as it is unlikely that a victim would be comfortable judging the kindness of an assailant!

It is perhaps worth mentioning that the procedure involving the default face may be of value to witnesses who, in spite of a good look at a suspect, cannot recall sufficient detail for an initial composite to be assembled; they may be denied the opportunity of constructing a composite in the UK. In this situation, construction could start from the default face using the procedure of Experiment 3. Alternatively, EvoFIT could be offered, given that a verbal description is not necessary but, either way, performance should be verified in further research, perhaps involving short target exposures to promote a weak memory and a sketchy description.

6.5. Composites in the real world

The current work sought a realistic design as far as possible in the laboratory. However, a potential limitation was the use of photographs and videos. Shapiro and Penrod (1986) found no change in participants' discrimination, and only a very minor change in bias, between face recognition studies involving live or videotaped stimuli and those involving photographs. Indeed, research on feature-based mugshot systems, also involving a significant recall component, have similarly found consistent results between photographs and a live target (Lee et al., 1998). Taken together, these data suggest that stimulus presentation is unlikely to affect composite quality. We note that videos and static photographs, while differing in their level of realism, did nonetheless produce consistent results: Experiments 1 to 3 found a consistent benefit for the serial interface, and Experiments 3 to 5 found a consistent effect of target encoding.

It is also possible that famous faces lessen the generalisation of the results, as suspects of crime are generally not celebrities. Comparing studies involving a broadly similar construction procedure from memory, similar naming levels are reported for those that have used celebrities (e.g. Brace et al., 2000, Frowd et al., 2004, 2005a) and those that have not (Davies et al., 2000, Bruce et al., 2002). A more important issue perhaps is that the distribution of composites in the real world is usually accompanied by additional information about the offender: age, build and Modus Operandi. While such information may provide cues to identification, it is known that

estimates of people tend to be unreliable (e.g. Cutshall and Yuille 1989), and thus the uncued naming task provides a measure of identification from the composite alone.

On a final note, the poor performance of standard PRO-fit after a 2 day delay is likely to raise concerns with law enforcement agencies. While the PRO-fit parallel interface was not the answer, EvoFIT results were very encouraging, with composites named more than twice as often as those from PRO-fit.

References

ACPO (2000). *National working practices in facial imaging*. Association of Chief Police Officers (Scotland) Working Group.

BOWER, G. H. and KARLIN, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology*, **103**, 751-757.

BRACE, N., PIKE, G. and KEMP, R. (2000). Investigating E-FIT using famous faces. In *Forensic psychology and law,* A. Czerederecka, T. Jaskiewicz-Obydzinska and J. Wojcikiewicz (Eds.), pp. 272-276 (Krakow: Institute of Forensic Research Publishers, 2000).

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.

CALDWELL, C. and JOHNSTON, V.S. (1991). Tracking a criminal suspect through "face-space" with a genetic algorithm, *Proceedings of the Fourth International Conference on Genetic Algorithms (*pp. 416-421). San Diego: Morgan Kaufmann.

CHRISTIE, D., DAVIES, G.M., SHEPHERD, J.W. and ELLIS, H.D. (1981). Evaluating a new computer-based system for face recall, *Law and Human Behaviour*, **2/3**, 209-218.

CLIFFORD, B.R. (2002, September). *The verbal overshadowing effect: In search of a chimera*. Paper presented at the 12th Conference of the European Association of Psychology and Law, Leuven, Belgium.

CUTLER, B. L., STOCKLEIN, C. J. and PENROD, S. D. (1988). An empirical examination of a computerized facial composite production system. *Forensic Reports*, **1**, 207-218.

CUTSHALL, J.L. and YUILLE, J.C. (1989). Field studies of eyewitness memory of actual crime scenes. In *Psychological methods in criminal investigation and evidence*, D.C. Raskin, (Ed.), pp. 97-124 (New York: Springer).

DAVIES, G.M. (1983). Forensic face recall: the role of visual and verbal information. In *Evaluating witness evidence*, S.M.A. Lloyd-Bostock and B.R. Clifford (Eds.), pp. 103-123 (New York: John Wiley and Sons Ltd).

DAVIES, G.M. and CHRISTIE, D. (1982). Face recall: an examination of some factors limiting composite production accuracy. *Journal of Applied Psychology*, **67**, 103-109.

DAVIES, G.M., ELLIS, H.G. and SHEPHERD, J. (1978). Face identification: The influence of delay upon accuracy of photofit construction, *Journal of Police Science and Administration*, **6**, 35-42.

DAVIES, G.M. and LITTLE, M. (1990). Drawing on memory: Exploring the expertise of a police artist. *Medical Science and the Law*, **30**, 345-354.

DAVIES, G.M., MILNE, A. and SHEPHERD, J. (1983). Searching for operator skills in face composite reproduction. *Journal of Police Science and Administration*, **11**, 405-9.

DAVIES, G.M., VAN DER WILLIK, P. and MORRISON, L.J. (2000). Facial Composite Production: A Comparison of Mechanical and Computer-Driven Systems. *Journal of Applied Psychology*, **85**, 119-124.

ELLIS, H.D., SHEPHERD, J.W. and DAVIES, G.M. (1980). The deterioration of verbal descriptions of faces over different delay intervals. *Journal of Police Science and Administration*, **8**, 101-106.

FROWD, C.D., CARSON, D., NESS, H., RICHARDSON, J., MORRISON, L., MCLANAGHAN, S. and HANCOCK, P.J.B. (2005a). A forensically valid comparison of facial composite systems. *Psychology, Crime and Law*, **11**, 33-52.

FROWD, C.D., CARSON, D., NESS, H., MCQUISTON, D., RICHARDSON, J., BALDWIN, H. and HANCOCK, P.J.B. (2005b). Contemporary Composite Techniques: the impact of a forensically-relevant target delay. *Legal and Criminological Psychology*, **10**, 63-81.

FROWD, C.D., HANCOCK, P.J.B. and CARSON, D. (2000, October). EvoFIT: A holistic, evolutionary face identification technique. *Presentation to the Association of Chief Police Officers*. Manchester.

FROWD, C.D., HANCOCK, P.J.B. and CARSON, D. (2004). EvoFIT: A holistic, evolutionary facial imaging technique for creating composites. *ACM Transactions on Applied Psychology (TAP)*, **1**, 1-21.

GEISELMAN, R.E., FISHER, R.P., MACKINNON, D.P. and HOLLAND, H.L. (1986). Eyewitness memory enhancement with the cognitive interview. *American Journal of Psychology*, **99**, 385-401.

GIBLING, F. and BENNETT, P. (1994). Artistic enhancement in the production of photofit likeness; an examination of its effectiveness in leading to suspect identification, *Psychology, Crime and Law*, **1**, 93-100.

GOFFREDSON, M.R. and POLAKOWSKI, M. (1995). Information retrieval: reconstructing faces. In *Psychology and Policing*, N. Brewer and C. Wilson (Eds.), pp. 101-117 (Hillsdale, New Jersey: LEA).

HANCOCK, P.J.B. and BRUCE, V. (2003). CRIME-VUs final report (EPSRC Grant GR/N09701). Unpublished.

HANCOCK, P.J.B., BRUCE, V. and BURTON, A.M. (1997). Testing principal component representations for faces. In *Proceedings of 4th Neural Computation and Psychology Workshop*, April 1997, J.A. Bullinaria, D.W.Glasspool and G. Houghton (Eds.), pp. 84-97 (London: Springer-Verlag).

HANCOCK, P.J.B., BRUCE, V. and BURTON, A.M. (2000). Recognition of unfamiliar faces. Trends in Cognitive Sciences, 4, 330-337.

KOEHN, C.E. and FISHER R.P. (1997). Constructing facial composites with the Mac-a-Mug Pro system. *Psychology, Crime and Law*, **3**, 215-224.

LAUGHERY, K.R., DUVAL, C. and WOGALTER, M.S. (1986). Dynamics of facial recall. In *Aspects of face processing,* Ellis, H.D., Jeeves, M.A., Newcombe, F. and Young, A. (Eds.), pp. 373-387 (Dordrecht: Martinus Nijhoff).

LAUGHERY, K.R., FESSLER, P.K., LENOROVITZ, D.R. and YOBLICK, D.A. (1974). Time delay and similarity effects in facial recognition. *Journal of Applied Psychology*, **59**, 490-496.

LEE, E., WHALEN, T., JOLLYMORE, G., READ, C. and SWAFFER, M. (1998). The effects of delay on the performance of computerized feature systems for identifying suspects. *Behaviour and Information Technology*, **17**, 294-300

MCQUISTON, D.E. (2003). *Verbal and visual processes in face recall: Assessing the role of perceptual expertise*. Unpublished doctoral dissertation, University of Texas at El Paso.

MEISSNER, C.A. and BRIGHAM, J.C. (2001). A meta-analysis of the verbal overshadowing effect in face identification, *Applied Cognitive Psychology*, **15**, 603-616.

MITCHELL, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: MIT.

RAKOVER, S.S. and CAHLON, B. (1996). To catch a thief with a recognition test: the model and some empirical results. *Cognitive Psychology*, **21**, 423-468.

SCHOOLER, J.W. and ENGSTLER-SCHOOLER, T.Y. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. *Cognitive Psychology*, **22**, 36-71.

SHEPHERD, J.W. (1983). Identification after long delays. In *Evaluating witness evidence*, S.M.A. Lloyd-Bostock and B.R. Clifford (Eds.), pp. 173–187 (Chichester: Wiley).

SHAPIRO, P. N. and PENROD, S.D. (1986). Meta-analysis of facial identification rates. *Psychological Bulletin*, **100**, 139-156.

TANAKA, J.W. and FARAH, M.J. (1993). Parts and wholes in face recognition, *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, **46A**, 225-245.

TANAKA, J.W. and SENGCO, J.A. (1997). Features and their configuration in face recognition. *Memory and Cognition*. **25**, 583-592.

TULVING, E. and THOMSON, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, **80**, 352-373.

WELLS, G.L. (1984). How adequate is human intuition for judging eyewitness testimony. In *Eyewitness testimony: Psychological perspectives*, G. R. Wells and E. F. Loftus (Eds.), pp. 256-272 (Cambridge, England: Cambridge University Press).

WELLS, G.L. and HRYCIW, B. (1984). Memory for faces: encoding and retrieval operations. *Memory and Cognition*, **12**, 338-344.

WELLS, G. L. and TURTLE, J.W. (1988). What is the best way to encode faces? In *Practical aspects of memory: current research and issues,* M.M. Gruneberg, P. Morris and R. Sykes (Eds.), pp. 163-168 (Chichester: Wiley).

WESTERMAN, D. L. and LARSEN, J. D. (1997). The verbal overshadowing effect: Evidence for a general shift in processing. *American Journal of Psychology*, **110**, 417-428.